

# 基于 SAM 法和 BP 神经网络的代谢综合长征风险、趋势预测和干预模型的构建

## 摘要

本文基于人体生物学数据,运用 SAM 算法、permutation 算法、BP 神经网络算法、logistic 回归分析、结合 MATLAB、SPSS 软件构建了样本相关基因挖掘模型、基于相互作用网络变化程度的动态模型、关键通路挖掘模型、代谢综合征风险预测模型,在基因表达量和关键通路方面分别对代谢综合征进行现态的患病等级预测和未来的患病风险预测,并根据预测情况给出具有针对性的干预措施。

针对问题一,本文采用 SAM 算法得到每个基因在相邻患病严重程度等级之间的差异值,同时结合假阳性检验选定阈值,筛选出大于阈值的基因为相关基因(如表 4 所示),将其录入到 KEGG 数据库中检索,得到包含相关基因的信号通路图,从信号图中挖掘出 66 条包含特征基因的相互作用网络,并利用模糊评价计算相互作用网络中每个基因的权值,进一步计算每条网络平均变化程度,用其作为每个阶段的度量,构建人体生命量化的动态模型。

针对问题二,利用 SPSS 软件通过 66 条相互作用网络对样本的患病等级进行多元 Logistic 回归分析,确定由问题一得到的 66 条相互作用网络对患病等级的影响程度,由总体的拟合效果以及似然比检验的显著性水平判断每条网络的影响程度,筛选出显著性小于 0.05 的作为关键通路,最终得到 30 条关键通路,更进一步的对剔除的网络再次进行多元 Logistic 回归分析以检验选取的合理性。

针对问题三,基于第一问所得的相关基因用 SAM 法最后筛选出两组特征基因、第二问所得的关键通路变化情况分别作为 3 个 BP 神经网络输入因子,选取部分数据进行训练,取剩余部分对训练好的 BP 神经网络进行测试,将未知患病严重程度的样本相关数据代入训练好的 3 个 BP 神经网络识别样本现状与未来患病风险,预测患病严重程度的变化趋势,给出如表 17 所示的个性化预防方案。

**关键字** SAM 算法 Permutation 算法 BP 神经网络 logistic 回归分析  
MATLAB SPSS

## 一、问题重述

人体作为一个复杂的非线性系统，其患代谢综合征的成因不仅与外部自然环境等相关，还与基因序列密不可分。每个人患代谢综合征的风险高低与人与人之间基因序列的差异有直接的关系，通过测量分析人体各类从内部到外部的因素明晰人体复杂系统的具体问题所在，构建早期趋势预测模型，从而完成个性化预防干预。试就一个具有较完整生物医学数据的人类群体建立数学模型，分析研究以下问题：

- (1) 分析 NCBI, EBI, DDBJ 等公开数据库中生物分子的相互作用和基因通路信息，构建人体生命量化模型；
- (2) 根据临床检测数据，找出代谢综合征关键通路的数据特征或相互作用网络；
- (3) 判断一个包含基因组、表观基因组、转录组、蛋白质组和代谢组的人类群体数据集患代谢综合征的风险，并对应找出造成他们代谢综合征风险的主要因素。

## 二、问题分析

### 2.1 背景分析

代谢综合征是一种能增加糖尿病、心脑血管疾病的患病率及其死亡风险的慢性疾病，虽然目前代谢综合征及其后续关联疾病的发病机制尚未明确，但相关研究已经表明它与遗传、环境、心理、生活方式、及年龄均有密不可分的关系，且是一个长期慢性发展的结果。根据美国新闻相关报导，目前全球已有四分之一的成年人患有代谢综合征，而患有代谢综合征的人患心脏病或中风的概率是正常人的 2-3 倍，因此对代谢综合征的预防和干预已经迫在眉睫。

本题是一个数据建模问题，主要是对生物体某个具体信息层次（如 DNA 序列，DNA 表达，基因相互作用等）之间的相互作用关系对患代谢综合征风险的研究。

### 2.2 问题一分析

在没有外在措施干预情况下，人体患代谢综合征的严重程度会随时间增加而递增，患者患病程度向下一程度等级变化时，其基因表达差异会相应变化。因此

先以批次一中 20073 个基因为样本，选择使用 SAM 法计算每一个基因在患病程度等级 I 与等级 II 之间的差异值，由于阈值选取不同，相关基因的筛选数目与假发现率也不同，通过假阳性检验选取合适的阈值来筛选相关基因，在查阅相关文献[2]的基础上我们假定患病程度共有四个等级（参考下文假设）。类似地我们再次使用同样的方法筛选出等级 II 和等级 III 之间、等级 III 和等级 IV 之间的相关基因，得到三组相关基因，同理，对批次二做同样的处理，两个批次将得到六组相关基因，为了能够包含每个组别的相关基因，对六组相关基因取并集，得到两个批次总的相关基因。

在参考 NCBI 和查阅了大量文献基础上，并对问题深入分析之后，考虑到了使用相互作用网络去构建人体生命量化的动态模型，即通过相互作用网络变化程度来衡量生命的每一阶段。为了挖掘相互作用网络，将得到的相关基因录入到 KEGG 中，进一步从 KEGG 中挖掘出相互作用网络。考虑到网络中基因与基因之间的复杂关系以及每个基因所起的作用不同，于是本文采用了模糊评价的方法衡量每个基因在相互作用网络的作用程度。通过变异基因的作用程度与网络中总基因的作用程度的比值衡量相互作用网络变化程度。最终我们构建了生命量化的动态模型。

## 2.3 问题二分析

关键通路旨在寻找问题一得到的相互作用网络中哪个网络跟代谢综合征患病程度的关系更为密切，故本文采用了 logistic 回归分析的方法寻找关键通路。将相互作用网络和患病等级录入到 spss 中进行回归分析，在模型拟合程度较好的基础上分析每一条网络对患病程度的影响是否显著，从而剔除影响不显著的相互作用网络。

为了验证我们剔除网络的合理性，我们对剔除掉的相互作用网络再次进行 logistic 回归分析。通过模型拟合程度以及似然比检验的显著性以证实我们所选取的关键通路是合理的。

## 2.4 问题三分析

代谢综合征的发展是一个长期的过程，最终诊断为代谢综合症已经是该长期过程的结果。我们认为通路的变化程度为人体最根本的内部因素，影响着每个人

罹患代谢综合疾病的风险高低，相关基因表达量则体现样本现阶段的患病严重程度。

在前两问的基础上，可得与代谢综合症有关的相关基因与关键通路，然而代谢综合征的具体分类标准还不明确，现有数据变量较多。BP 神经网络有较强的处理不严密知识和定性特征变故的能力<sup>[1]</sup>，故在预测方面我们采用这种人工智能技术。针对变量较多的问题，对每批次的三个相关基因集合（批次一  $U^A$ 、 $U^B$ 、 $U^C$  和批次二  $U^D$ 、 $U^E$ 、 $U^F$ ）再进行 SAM 法筛选，分批次的得到随患病等级变化差异较大的特征基因。

综合由特征基因表达量预测的患病严重程度与关键通路变化程度预测的患病严重程度分析样本代谢综合征得发展趋势，并给出相关建议。

### 三、模型假设

- （1）假设基因表达量是长期作用结果，不考虑近期变异。
- （2）假设样本变异情况只考虑编码氨基酸的影响。
- （3）变异除 silent 变异类型以外，其他变异均对相关基因作用网络产生影响。
- （4）假设 I A 和 I B 统称为 I，II A 和 II B 统称为 II，III A 和 III B 统称为 III，IV A 和 IV B 统称为 IV。<sup>[2]</sup>
- （5）假设样本只涉及代谢综合征，不涉及其他疾病。

### 四、符号说明

$X1(i)$	第 i 个基因在第一类样本中的均值
$X2(i)$	第 i 个基因在第二类样本中的均值
$n1$	第一类样本容量
$n2$	第二类样本容量
$s(i)$	第 i 个基因的标准偏差
$d(i)$	第 i 个基因的统计量值
$dE(i)$	第 i 个基因的统计量值的均值
$D$	基因差异值

$U^A$	第一批 I 与 II 表达差异性基因
$U^B$	第一批 II 与 III 表达差异性基因
$U^C$	第一批 III 与 IV 表达差异性基因
$U^D$	第二批 I 与 II 表达差异性基因
$U^E$	第二批 II 与 III 表达差异性基因
$U^F$	第二批 III 与 IV 表达差异性基因
$fdr$	假阳性检验所得的假发现率
$\omega_{ni}'$	第 n 个样本所包含变异基因的第 i 条相互作用网络的权值
$y_{ni}$	第 n 个样本第 i 个相互作用网络的变化程度
$m1$	Stage I 阶段总人数
$m2$	Stage II 阶段总人数
$m3$	Stage III 阶段总人数
$m4$	Stage IV 阶段总人数
$Y_1$	第 I 阶段相关作用网络平均变化程度
$Y_2$	第 II 阶段相关作用网络平均变化程度
$Y_3$	第 III 阶段相关作用网络平均变化程度
$Y_4$	第 IV 阶段相关作用网络平均变化程度

## 五、模型建立与求解

### 5.1 问题一的模型建立与求解

#### 5.1.1 生命量化动态模型的建立

本文的生命量化动态模型是基于相互作用网络的变化程度构建的,而构建这个动态模型需要挖掘出相互作用网络和计算网络变化程度。挖掘相互作用网络需从两万多个基因中找到和代谢综合征相关的相关基因,本文采用 SAM 法<sup>[3,4]</sup>与

假阳性检验筛选相关基因并利用模糊评价的方法计算网络变化程度，具体步骤如下：

第一步：**SAM** 法与假阳性检验筛选相关基因。分两批次进行，对每批次两个相邻之间的等级使用一次 **SAM** 法与假阳性检验筛选出之一组的相关基因，最终求其并集；

第二步:相互作用网络的挖掘。参考 **NCBI** 和 **KEGG** 数据库并结合相关基因找到信号通路图，从图中挖掘相互作用网络；

第三步：模糊评价其变化程度。利用模糊评价的方式以经过某一基因的作用线作为衡量这一基因在此通路图中的权值，结合样本变异基因计算网络变化程度；

第四步：建立动态模型。以患病等级为生命的每一个阶段，以相互作用网络的平均变化程度作为每一阶段的量度。

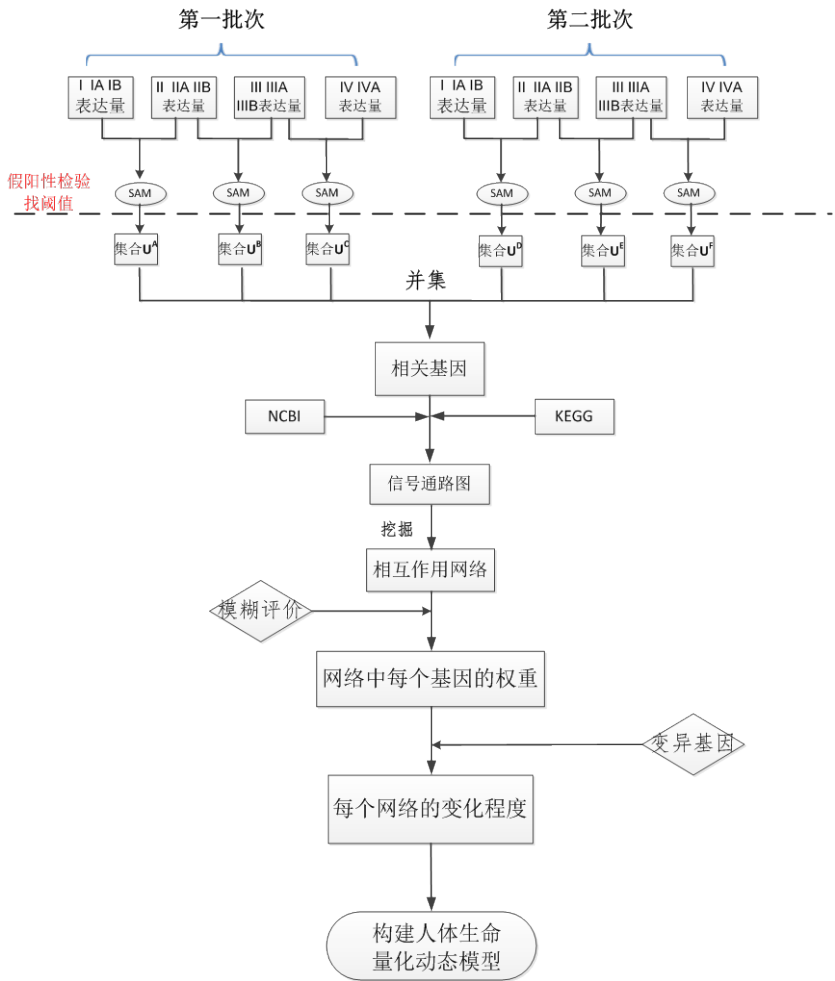


图 1 为其流程图

### (1)SAM 法与假阳性检验筛选相关基因

①从第一批样本表达文件中选出患病程度等级为 I 和等级为 II 的基因,分别记为 $n_1$ 和 $n_2$ ,计算每个基因的标准偏差,记为 $s(i)$ :

$$s(i) = \sqrt{a \left[ \sum_{n=1}^{n_1} (x_n(i) - x_1(i))^2 + \sum_{m=1}^{n_2} (x_m(i) - x_2(i))^2 \right]} \quad (1)$$

其中

$$a = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2} \quad (2)$$

计算每个基因的统计量值,记为 $d(i)$ :

$$d(i) = \frac{x_1(i) - x_2(i)}{s(i) + s_0} \quad (3)$$

$s_0$ 为变异系数最小时对应的 $s(i)$ ,通常为很小的正数, $d(i)$ 的变异系数为 $s(i)$ 的函数。

②采用 permutation 算法对样本进行全排列,再以前 $n_1$ 个基因作为第一类样本,后面 $n_2$ 个基因作为第二类样本,重复过程(1)计算新样本中各个基因的统计量值。

③将过程②进行  $n$  次计算,对第  $p$  次 permutation 算法第  $i$  个基因的统计量值,记为 $d_p(i)$ :

求统计量值的均值,记为 $d_E(i)$ :

$$d_E(i) = \frac{\sum_{p=1}^n d_p(i)}{n} \quad (4)$$

计算基因差异值,记为  $D$ :

$$D(i) = |d(i) - d_E(i)| \quad (5)$$

将所得的 $d(i)$ 、 $d_p(i)$ 、 $n$ 进行假发现率检验:

$$fdr = \frac{\sum_{p=1}^n \#\{i | |d_p(i)| > |d(i)|\}}{n \sum \#\{i | D(i) > t\}} \quad (6)$$

其中  $t$  为阈值,  $\#$ 表示对集合求元素个数。

④根据假发现率随阈值  $t$  的变化情况,选取合适的阈值  $t$  使  $\text{fdr} < 0.005$ ,筛选出  $D$  大于阈值  $t$  的基因为差异表达基因。

⑤重复①—④过程,筛选出第一批样本表达文件中患病程度等级为 II 和等级为 III 之间、等级为 III 和等级为 IV 之间的差异表达基因,对所选出的差异表达基因取并集,得到相关基因。

## (2)模糊评价与相互作用网络的挖掘

①提取相关基因的 ID 号通过 KEGG 数据库的信号通路图寻找相关基因的相互作用网络。根据所选取相关基因的表达量与患病程度等级之间的数字相关即相邻患病等级之间相关基因的表达量有较大的差异值,提取的相互作用网络中包括相关基因与每条相互作用线(经过相关基因)上紧邻相关基因的两个有关基因。

②在一个相互作用网络中,由于每一个基因的作用程度不同,根据网络中的基因所经过的流程个数,利用模糊评价的方法评判网络中每个基因所起到的作用程度。

③由上一步可以得到每一个样本包含相关基因的相互作用网络,利用模糊评价得到相互作用网络中每一个基因的影响程度为  $\omega_{ij}$ 。

## (3)相互作用网络的变化程度

④样本每条通路中的基因可能会发生变异,跟据模型假设,剔除发生 silent 变异类型的基因,提取剩余变异类型的基因,这些基因在相互作用网络中的影响程度为  $\omega_{ik} (k < j)$ ,可以得到每个样本的每一条通路的权值。

第  $n$  个样本所包含变异基因的第一条相互作用网络的权值:

$$\omega_{n1}' = \omega_{n11} + \omega_{n12} + \omega_{n13} + \cdots \omega_{n1k} \quad (7)$$

第  $n$  个样本所包含变异基因的第二条相互作用网络的权值:

$$\omega_{n2}' = \omega_{n21} + \omega_{n22} + \omega_{n23} + \cdots \omega_{n2k} \quad (8)$$

.....

第  $n$  个样本所包含变异基因的第  $i$  条相互作用网络的权值:

$$\omega_{ni}' = \omega_{ni1} + \omega_{ni2} + \omega_{ni3} + \cdots \omega_{nik} \quad (9)$$

第  $n$  个样本第一个相互作用网络的变化程度:



$$y_{n1} = \frac{\omega_{n1}'}{\omega_{n11} + \omega_{n12} + \omega_{n13} + \cdots \omega_{n1j}} \quad (10)$$

第 n 个样本第二个相互作用网络的变化程度:

$$y_{n2} = \frac{\omega_{n2}'}{\omega_{n21} + \omega_{n22} + \omega_{n23} + \cdots \omega_{n2j}} \quad (11)$$

.....

第 n 个样本第 i 个相互作用网络的变化程度

$$y_{ni} = \frac{\omega_{ni}'}{\omega_{ni1} + \omega_{ni2} + \omega_{ni3} + \cdots \omega_{nij}} \quad (12)$$

第 n 个样本共 i 个相互作用网络的平均变化程度

$$y_n = \frac{y_{n1} + y_{n2} + y_{n3} + \cdots + y_{ni}}{i}$$

#### (4) 基于相互作用网络的生命量化动态模型

由于代谢综合征为慢性疾病,是长期发展的结果,在不加干预的情况患病程度的等级也会随着时间的变化而逐步加深,因此将四个不同的患病等级作为人体生命的不同的阶段,将相互作用网络的变化程度作为人体生命不阶段的度量。

第 I 阶段相关作用网络平均变化程度:

$$Y_1 = \sum_{n=1}^{m1} y_n / m1 \quad (13)$$

第 II 阶段相关作用网络平均变化程度:

$$Y_2 = \sum_{n=m1+1}^{m2+m1} y_n / m2 \quad (14)$$

第 III 阶段相关作用网络平均变化程度:

$$Y_3 = \sum_{n=m1+m2+1}^{m3+m2+m1} y_n / m3 \quad (15)$$

第 IV 阶段相关作用网络平均变化程度:

$$Y_4 = \sum_{n=m1+m2+m3+1}^{m1+m2+m3+m4} y_n / m4 \quad (16)$$

### 5.1.2 生命量化动态模型求解

#### (1) SAM 法与假阳性检验筛选相关基因

通过附录一程序得到第一批样本表达文件中患病程度等级为 I 和等级为 II 之间 20073 个基因的差异值排名, 程序附录二为假阳性检验,得到如下表:

	A	B	C	D	E	F	G	H	I	J	K
1	阈值	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
2	假样率	0.403538	0.346865	0.290681	0.232023	0.179373	0.135645	0.100939	0.070442	0.048442	0.033808
3	阈值	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
4	假样率	0.022721	0.014592	0.009125	0.00558	0.003496	0.002181	0.001166	0.000549	3.02E-06	8.06E-09

根据假发现率小于 0.005 和代谢综合征的复杂性,这里取阈值为 1.4, 最终得到差异基因个数为 130 个,记为集合  $U^A$ ,如表 1 所示(由于篇幅问题只展示部分):

	A	B	C	D	E	F
1	基因名	差异值	基因名	差异值	基因名	差异值
2	LUZP4   51213	2.2590941	RAB5A   5868	1.654893404	FOXP1   27086	1.55047
3	TCN1   6947	2.07132839	KCNK6   9424	1.649774708	FGF16   8823	1.54904
4	RNF32   140545	2.055729148	BDKRB2   624	1.649631858	MT01   25821	1.54821
5	NET1   10276	1.947525809	NAGPA   51172	1.623829687	ZNF606   80095	1.54687
6	ANXA2   302	1.878883845	PPL   5493	1.623784836	PILRB   29990	1.54409
7	ZNF862   64364	1.866318289	FLJ44635   392490	1.61767537	FAM108C1   58489	1.54105
8	MAPK6   5597	1.840243169	MYD88   4615	1.606240743	ANXA2P1   303	1.53816
9	MESDC2   23184	1.839012023	SHISA5   51246	1.603694844	KISS1   3814	1.53704
10	TBX18   9096	1.835107641	HSP90AB4P   664618	1.60168993	CLP1   10978	1.53446
11	PPP3CA   5530	1.82625333	ZNF329   79673	1.600017173	CATSPERG   57828	1.53257
12	CAMP   820	1.80510318	LEO1   123169	1.593683966	LOC200030   200030	1.53224
13	ZNF808   38855	1.798707094	WASH2P   375260	1.593184472	PTPLB   201562	1.53203
14	GJA1   2697	1.767197414	CEP290   80184	1.591616239	C2CD4A   145741	1.52749
15	SERPINB8   527	1.759731544	TSHR   7253	1.586746487	SEC31A   22872	1.52658
16	PDGFRL   5157	1.734333172	SEMA3C   10512	1.577140318	SLMAP   7871	1.52309
17	LOC150786   15	1.73198035	IDI1   3422	1.576427485	REPIN1   29803	1.52116
18	LOC158572   15	1.728207321	POLG2   11232	1.573411479	PLCB3   5331	1.51947
19	LENG9   94059	1.721350154	PDE6A   5145	1.571883689	ZNF665   79788	1.51756
20	LDLR   3949	1.71551537	DOCK10   55619	1.570755646	ZNF382   84911	1.51346

表 1 I 与 II 表达差异性基因

同理,得到第一批样本表达文件患病等级为 II 与患病等级为 III 的假阳性检验表:

	A	B	C	D	E	F	G	H	I	J	K
1	阈值	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
2	假样率	0.440001	0.383926	0.329814	0.27891	0.220892	0.165845	0.127117	0.09051	0.064397	0.042439
3	阈值	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
4	假样率	0.0279	0.017726	0.010424	0.006568	0.004253	0.002731	0.002415	0.002578	0.004036	65535

最终得到 38 个差异基因,记为集合  $U^B$ ,如表 2 所示:

	A	B	C	D	E	F	G
1	基因名	差异值	基因名	差异值	基因名	差异值	
2	MOBK13   25843	1.932182953	MESDC2   23184	1.659450706	MAMDC2   256691	1.576613	
3	SELS   55829	1.878592978	AP1G1   164	1.654144359	ENTPD6   955	1.574414	
4	ZNF510   22869	1.873902529	KAL1   3730	1.641514755	IDI1   3422	1.568242	
5	MON1B   22879	1.839869902	CRYBA1   1411	1.62877005	SCOC   60592	1.567896	
6	NFKBIE   4794	1.830194956	FUT10   84750	1.628316483	RFPL4A   342931	1.563604	
7	PDP2   57546	1.823619263	SLC16A3   9123	1.624386001	PLEKHB1   58473	1.551908	
8	SLC25A10   1468	1.801658593	CXCL14   9547	1.618577293	ANKFN1   162282	1.549664	
9	C1RL   51279	1.764210127	HTRA2   27429	1.613985643	SAP30BP   29115	1.548143	
10	MTUS1   57509	1.742950068	ACVR2A   92	1.613464141	INTS10   55174	1.546806	
11	COPB1   1315	1.741241417	ELFN1   392617	1.607210466	C4orf41   60684	1.544389	
12	AZI2   64343	1.697614005	RNF141   50862	1.606793431	BCL7B   9275	1.542193	
13	TAF5L   27097	1.681906394	SLC10A7   84068	1.591935258	MRPS24   64951	1.540328	
14	UBC   7316	1.677997581	C15orf38   348110	1.591259165	CSTF2T   23283	1.528091	
15	PMS2   5395	1.676100668	ZNF45   7596	1.588399195	NAF1   92345	1.52161	
16	DCTD   1635	1.671193434	IL20RA   53832	1.584956933	FTSJ1   55783	1.518833	
17	RPS12   6206	1.669387728	STAT6   6778	1.584354098	IRF5   3663	1.512391	

表 2 II 与 III 表达差异性基因

同理,得到第一批样本表达文件患病等级为Ⅲ与患病等级为Ⅳ的假阳性检验表:

	A	B	C	D	E	F	G	H	I	J	K
1	阈值	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
2	假样率	0.453933	0.392357	0.333677	0.277487	0.208546	0.152426	0.11195	0.077634	0.051589	0.034511
3	阈值	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
4	假样率	0.021896	0.014358	0.008528	0.005625	0.003465	0.002104	0.001566	0.000507	1.35E-06	8.11E-09
5											

最终得到 46 个差异基因,记为集合  $U^C$ ,如表 3 所示:

	A	B	C	D	E	F
1	基因名	差异值	基因名	差异值	基因名	差异值
2	OBSCN   84033	2.01478997	PLCZ1   89865	1.629799813	GLRB   2743	1.532148009
3	MYNN   55892	1.95765311	ZAK   51776	1.629482747	EIF4EBP1   19	1.53190034
4	SCML1   6322	1.93861338	MFS7   84179	1.619361323	PPME1   51400	1.529993327
5	TNFSF13   874	1.82670125	NOP56   10528	1.619201885	SDHAP1   2558	1.529556558
6	EIF1B   10289	1.78824838	STOML2   3096	1.618257086	FARP1   10160	1.525497991
7	KCTD14   6598	1.75296308	PTTG1   9232	1.614509093	SH3GLB1   511	1.523340681
8	SLC43A2   124	1.71443723	SLC25A6   295	1.600707232	GTF2IP1   297	1.520931935
9	GTF2E2   2961	1.70424532	SEC22A   2698	1.598478643	STRC   161497	1.519740504
10	C17orf63   55	1.70200487	ATF4   468	1.584914682	ZMYM3   9203	1.518487375
11	HIST1H3G   83	1.69267857	PTPN1   5770	1.58213427	CCDC47   5700	1.510907266
12	MKKS   8195	1.68099382	RPL36   25875	1.577023045	RBM18   92400	1.509915413
13	SNORA40   677	1.66259171	DIMT1L   2729	1.56202163	GZF1   64412	1.503594223
14	DDX10   1662	1.65630342	GLRX2   51022	1.555642749	DNAJC28   549	1.503413762
15	GPC4   2239	1.65210093	METTTL4   6486	1.550140412	SPOP   8405	1.503089582
16	C19orf48   84	1.64779899				

表 3 III与Ⅳ表达差异性基因

类似于第一批样本表达文件的处理方法,可以得到第二批样本表达文件中患病程度等级为Ⅰ和等级为Ⅱ的差异基因  $U^D$ ,患病程度等级为Ⅱ和等级为Ⅲ的差异基因  $U^E$ ,患病程度等级为Ⅲ和等级为Ⅳ的差异基因  $U^F$ 。(差异基因  $U^D$ 、 $U^E$ 、 $U^F$  数据和对应的阈值表见附录,篇幅原因这里不再展示)

接着对所得六组集合进行并运算,即  $U^A \cup U^B \cup U^C \cup U^D \cup U^E \cup U^F$ ,得到如表 4 所示 1818 个相关基因。(限于篇幅,这里只做部分展示)

	A	B	C
1	相关基因	相关基因	相关基因
2	ACOT11   26027	IRF5   3663	RNF115   27246
3	ACVR2A   92	KAL1   3730	RNF141   50862
4	ANKFN1   162282	KCNK6   9424	RNF32   140545
5	ANXA2P1   303	KCTD14   65987	RNPC3   55599
6	ANXA2P2   304	KISS1   3814	RPL36   25873
7	ANXA2   302	LDLR   3949	RPS12   6206
8	AP1G1   164	LENG9   94059	SAP30BP   29115
9	ATF4   468	LEO1   123169	SC4MOL   6307
10	AZI2   64343	LOC100270710   100270710	SCML1   6322
11	BCL7B   9275	LOC150786   150786	SCOC   60592
12	BDKRB2   624	LOC158572   158572	SDHAP1   255812
13	C15orf38   348110	LOC200030   200030	SEC22A   26984
14	C17orf63   55731	LOC723972   723972	SEC31A   22872
15	C19orf48   84798	LSR   51599	SELS   55829
16	C1RL   51279	LUZP4   51213	SEMA3C   10512
17	C2CD4A   145741	MAMDC2   256691	SERPINB8   5271
18	C4orf41   60684	MAPK6   5597	SH3GLB1   51100
19	CAMP   820	MDM1   56890	SHISA5   51246
20	CATSPERG   57828	MESDC1   59274	SLC10A7   84068
21	CCDC47   57003	MESDC2   23184	SLC16A3   9123
22	CDSN   1041	METTL4   64863	SLC25A10   1468
23	CEP290   80184	MFSD7   84179	SLC25A6   293
24	CHPF   79586	MKKS   8195	SLC38A2   54407
25	CLP1   10978	MOBK13   25843	SLC43A2   124935
26	COPB1   1315	MON1B   22879	SI MAP   7871

表 4 相关基因总集合

## (2)模糊评价与相互作用网络挖掘求解

将总相关基因带入到 KEGG 数据库中检索,得到包含相关基因的信号源通路图,从信号图中挖掘包含相关基因的相互作用网络,并利用模糊评价计算相互作用网络中每个基因的权值,图 2 为其中的一个信号通路图:

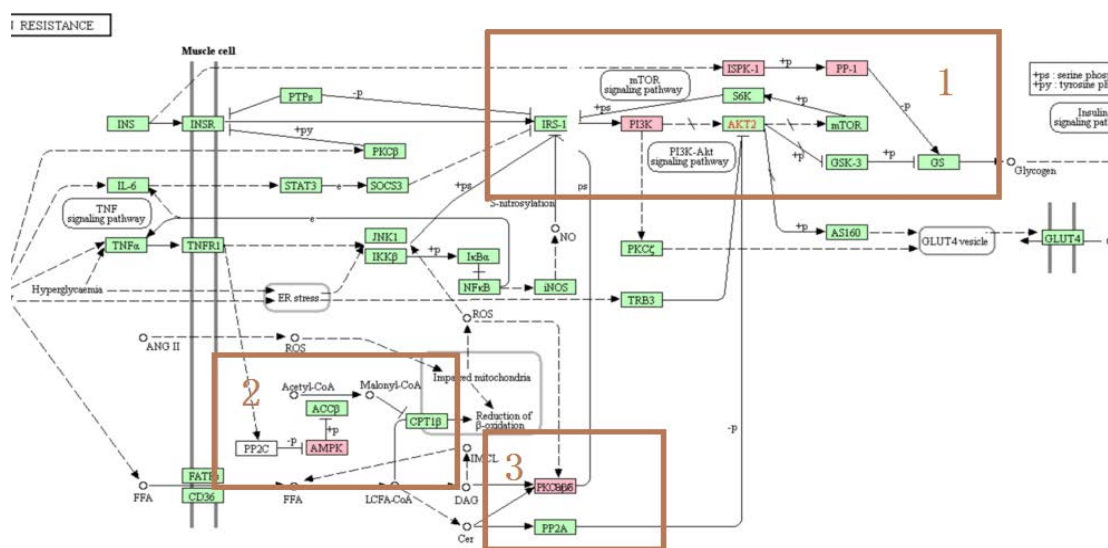


图 2 信号通路图

从中提取三个相互作用网络,并得到每个相互作用网络中相关基因的权值,三个网络的每个基因的权值如表 5 所示:

网络31		网络32		网络33	
基因	$\omega_{n30j}$	基因	$\omega_{n31j}$	基因	$\omega_{n32j}$
INS	1	IRS1	5	PRKCB	4
RPS6KA1	2	PIK3CA	3		
RPS6KA3	2	PIK3R1	3		
RPS6KA2	2	PIK3CG	3		
RPS6KA6	2	PIK3R5	3		
PPP1R3A	2	PIK3CB	3		
PPP1R3B	2	PIK3R2	3		
PPP1R3D	2	AKT2	4		
		PRKCZ	2		
		PRKAG3	2		
		ACACB	2		
		PRKAA2	2		
		PRKAA1	2		
		PRKAB1	2		

表 5 相关基因权值表

重复以上的挖掘相互作用网络的方式,在 42 个信号通路图中,我们将得到 66 个相互作用网络,如表 6 所示:

	A	B	C	D	E	F	G	H	I	J	K	L
1	网络1		网络2		网络3		网络4		网络5		网络6	
2	基因	$\omega_{n1j}$		$\omega_{n2j}$		$\omega_{n3j}$		$\omega_{n4j}$		$\omega_{n5j}$		$\omega_{n6j}$
3	PRKAA2	2	CSF1	1	LPAR4	2	XCR1	2	CCL4L2	1	IL4R	1
4	PRKAA1	2	EGFR	1	LPAR3	2	XCL2	1	CCL27	1	IL4	1
5	DDIT4	1	EGF	1	LPAR1	2	XCL1	1	CCL14	1		
6	TSC1	4	FGF6	1	F2R	2	PPBP	1	CCL15	2		
7	TSC2	4	FGF13	1	CHRM2	2	PF4	1	CCL2	1		
8	RHEB	2	FGF7	1	GNB2	2	CXCR6	1	CCL25	1		
9	MTOR	3	FGF23	1	GNG12	2	CXCR4	1	CCL3	2		
10	RPTOR	3	FGF5	1	PIK3CG	2	CXCR3	3	CCR1	7		
11			FGF4	1	PIK3R5	2	CXCR1	1	CCR3	9		
12			FGF14	1	RPS6KB1	2	CXCL6	2	CCR2	4		
13							CXCL14	1	CCR7	2		
14			...		...		CXCL11	2	CCR4	2		
15			ATF6B	2	AKT3	22	CX3CR1	1				
16			ATF2	2	AKT1	22						
17			MCL1	2	AKT2	22						
18	网络7		网络8		网络9		网络10		...		网络66	

	A	B	C	D	E	F	G	H	I	J	K	L
17			MCL1		2 AKT2	22						
18	网络7		网络8		网络9		网络10		...		网络66	
19	基因	$\omega_{n7j}$	基因	$\omega_{n8j}$	基因	$\omega_{n9j}$	基因	$\omega_{n10j}$			基因	$\omega_{n66j}$
20	IFNA16	1	IL22	1	INHBE	4	CACNA1A	1			CSGALNAC	2
21	IFNA21	1	IL19	1	INHBA	4	CACNA1B	1			CSGALNAC	2
22	IFNAR2	5	IL17B	1	INHBC	4	CACNA1C	1			CHSY3	2
23	IFNA17	1	IL17A	1	ACVR2A	1	CACNA1D	1			CHPF	2
24	IFNGR1	1	IL10RA	1	ACVR1B	1	CACNA1E	1			CHPF2	2
25	IFNA6	1			ACVR1	1	CACNA1F	1				
26	IFNA2	1					CACNA1G	1				
27	IFNGR2	1					CACNA1H	1				
28	IFNA14	1					CACNA1I	1				
29	IFNG	2					CACNA1S	1				
30	IFNK	2					CACNA2D1	1				
31	IFNA8	1					CACNA2D2	1				
32							CACNA2D3	1				
33							CACNA2D4	1				
34												
35							...					
36							RRAS2	7				
37							STK4	1				
38							ZAK	1				
39												
40												

表 6 相互作用网络

### (3)基于相互网络变化程度的动态模型求解

利用附录程序三寻找每个样本的变异基因,代入公式(7)、(8)、(9)计算每个样本包含变异基因的相互作用网络的权值,得到 $\omega_{n1}'$ 、 $\omega_{n2}'$ 、 $\omega_{n66}'$ , $n$  为第  $n$  个样本,共 66 个相互作用网络,根据(10)、(11)、(12)公式计算每个样本的 66 条网络的变化程度即  $y_{n1}$ 、 $y_{n2}$ 、...  $y_{n66}$ , $n$  为第  $n$  个样本,如表 7 所示(限于篇幅,只展示部分数据):



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	#Sample	网络1	网络2	网络3	网络4	网络5	网络6	网络7	网络8	网络9	网络10	网络11	网络12
2	82	0	0.166667	0	0	0	0	0	0	0	0	0.388889	0
3	83	0	0	0	0	0.111111	0	0	0	0	0	0.055556	0
4	84	0.333333	0.055556	0.083333	0	0	0	0	0	0	0	0	0
5	85	0	0	0	0	0	0	0	0	0	0	0	0
6	88	0	0	0	0	0	0	0	0	0	0	0	0
7	94	0	0	0	0	0	0	0	0	0	0	0	0
8	98	0	0	0	0	0	0	0	0	0	0	0	0
9	99	0.444444	0.444444	0.083333	0.333333	0	0	1	1	0	0.357143	0.777778	0
10	102	0	0.166667	0	0	0	0	0.166667	0	1	0.071429	0.166667	0
11	103	0	0.277778	0	0.333333	0	0	0	0	0	0	0.222222	0
12	104	0	0.083333	0	1	0	0	0	0	0.25	0	0.277778	0
13	107	0	0.083333	0	0	0	0	0	0	0	0	0.388889	0
14	108	0.333333	0.138889	0	0	0	0	0	1	0.25	0.5	0.166667	0.5
15	109	0	0.388889	0	0	0	0	0	0	0	0.285714	0	0
16	110	0	0.444444	0.166667	0	0	0	0.166667	0	0	0.642857	0.777778	0.5
17	111	0	0.722222	0.083333	0	0	0	0	0	0	0.5	0	0.5
18	116	0	0.166667	0	0.666667	0	0	0.166667	0	0	0.071429	0.055556	0
19	117	0	0.055556	0	0	0	0	0	0	0	0	0	0
20	118	0	0.194444	0.083333	0	0	0	0	0	0	0.285714	0	0.5
21	119	0.777778	0.888889	0.166667	0.666667	0	1	0	0	0	0.857143	0.5	0.5
22	144	0	0.194444	0	0	0	0	0	0	0	0	0	0
23	159	0	0.055556	0	0	0	0	0	0	0	0	0	0
24	165	0.333333	0	0	0	0	0	0	0	0	0	0	0
25	166	0	0.027778	0	0	0	0	0	0	0	0	0.055556	0
26	168	0	0	0	0	0	0	0	0	0	0.071429	0	0
27	171	0	0.083333	0	0	0	0	0	0	0	0.071429	0	0
28	173	0	0	0	0	0	0	0	0	0	0	0	0
29	174	0	0.055556	0	0	0	0	0	0	0	0	0	0
30	178	0	0.027778	0	0.666667	0	0	0	0	0	0	0	0

	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO
1	网络51	网络52	网络53	网络54	网络55	网络56	网络57	网络58	网络59	网络60	网络61	网络62	网络63	网络64	网络65	网络66
2	0.444444	0	0	0	0.333333	0.173913	0	0	0	0.5	0	0.5	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0.130435	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	0.833333	0.608696	0.333333	0.521739	0	0	0	0.5	0	0	0	0	0	0
10	0.222222	0	0.25	0.26087	0	0	0.181818	0	0	0	0	0	0	0	0	0
11	0.666667	0	0.25	0.652174	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0.166667	0.130435	0.333333	0.173913	0	0	0	0.5	0	0.5	0	0	0	0
14	0.666667	0	0.5	0.130435	0	0.217391	0	0	0	0	0	0	0	0	0	1
15	0.222222	0	0	0.217391	0	0.26087	0	0.285714	0	0	0	0	0	0	0	0
16	0	0.333333	0.583333	0.217391	0.666667	0.565217	0	0	0	1	0	1	0	0	0	0
17	0.555556	0	0.5	0.26087	0	0.086957	0.181818	0	0	0	0	0	0	0	0	0
18	0.444444	0	0	0.26087	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	1	0	0	0.086957	0	0	0	0	0	0	0	0	0	0
21	0	0	0.583333	0.217391	0.333333	0.782609	0	0.285714	0	0.5	0	0.5	1	0	0	1
22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0.130435	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0.086957	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0.217391	0	0	0	0	0	0	0	0	0	0
27	0	0	0.333333	0.217391	0	0.086957	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0.130435	0	0.173913	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

表 7 网络变化程度

得到每个样本的 66 条相互作用网络变化程度之后,由于两批次患病等级为 I 的人数为 45 人即  $m_1=4$ , 患病等级为 II 的人数为 34 即  $m_2=34$ , 患病等级为 III 的

人数为 28 即  $m_3=28$ , 患病等级为 IV 的人数为 28 即  $m_4=28$ , 代入公式(13)、(14)、(15)、(16)得到两批次患病等级为 I、II、III、IV 的平均变化程度如表 8 所示:  
(限于篇幅, 只展示部分数据)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	#Sample	平均影响程度	患病等级		#Sample	平均影响程度	患病等级		#Sample	平均影响程度	患病等级		#Sample	平均影响程度	患病等级	
2	217	0.136775036	1		142	0.180397	2		114	0.353913	3		150	0.211234	4	
3	208	0.133139653	1		135	0.159775	2		213	0.272162	3		202	0.210577	4	
4	205	0.128737523	1		138	0.145472	2		226	0.191326	3		79	0.182199	4	
5	225	0.122912807	1		115	0.14356	2		139	0.135212	3		151	0.1633	4	
6	102	0.119705767	1		100	0.130195	2		132	0.118795	3		133	0.136286	4	
7	82	0.109307169	1		152	0.127705	2		210	0.102569	3		134	0.130482	4	
8	107	0.106248837	1		184	0.109478	2		214	0.096694	3		154	0.120661	4	
9	103	0.105678224	1		157	0.083052	2		141	0.089526	3		76	0.11012	4	
10	221	0.102638829	1		75	0.079841	2		87	0.065192	3		149	0.10041	4	
11	227	0.097128713	1		212	0.074295	2		106	0.056928	3		192	0.073102	4	
12	228	0.089347086	1		209	0.070522	2		96	0.051768	3		181	0.040303	4	
13	224	0.080109515	1		78	0.070301	2		167	0.042495	3		182	0.038212	4	
14	118	0.073522457	1		206	0.060535	2		112	0.028794	3		160	0.036201	4	
15	197	0.069556852	1		120	0.059462	2		190	0.027057	3	0.063735	143	0.034342	4	0.064924
16	229	0.060472621	1		216	0.055617	2	0.05905	140	0.024531	3		136	0.031618	4	
17	178	0.055976431	1		158	0.052981	2		164	0.020454	3		161	0.026595	4	
18	222	0.055973737	1		145	0.051973	2		86	0.01881	3		172	0.024192	4	
19	220	0.053643725	1		147	0.049722	2		137	0.017429	3		156	0.020004	4	
20	116	0.05345109	1		219	0.046961	2		113	0.015552	3		187	0.019837	4	
21	117	0.052588207	1		148	0.042673	2		177	0.015023	3		176	0.019731	4	
22	104	0.051683502	1		155	0.038991	2		194	0.011271	3		80	0.018496	4	
23	203	0.043991147	1		193	0.03002	2		191	0.007576	3		200	0.014603	4	
24	109	0.038760201	1	0.04869	146	0.027893	2		170	0.007155	3		195	0.013076	4	
25	171	0.026472359	1		186	0.026723	2		163	0.004377	3		74	0.010859	4	
26	215	0.026010101	1		105	0.01964	2		175	0.004377	3		73	0.008507	4	
27	84	0.024400211	1		185	0.017556	2		211	0.002225	3		162	0.008165	4	
28	223	0.020789221	1		77	0.012714	2		93	0.002104	3		81	0.007728	4	
29	144	0.019692539	1		196	0.010732	2		95	0.001263	3		92	0.007027	4	
30	201	0.010106550	1		180	0.007540	2									

表 8 平均变化程度

将患病等级 I、II、III、IV 作为人体生命变化的四个阶段,用相互作用网络的平均变化程度作为每个阶段的度量构建人体生命量化的动态模型。 如图 3 所示:



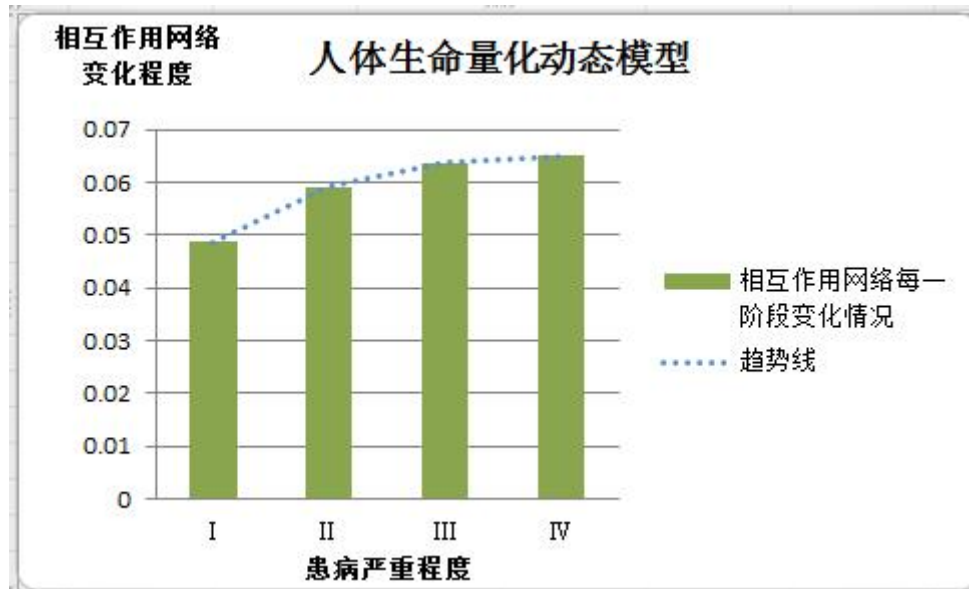


图3 人体生命量化动态模型

## 5.2 问题二的模型建立与求解

### 5.2.1 关键通路模型建立

问题一通过 SAM 法与假阳性检验筛选出相关基因,紧接着根据 KEGG 数据库查找到相关基因的通路图,从通路图中得到 66 条相互作用网络,问题二是模型将从这 66 条相互作用网络中筛选出主要网络作为关键通路,利用多元 Logistic 回归分析确定这 66 条网络对患病等级的影响程度。通过 SPSS 多元 Logistic 回归分析可以得到模型拟合信息和拟合优度,根据模型拟合信息和拟合优度来判断模型能否更好地拟合数据,进一步根据似然比检验中的似然值、卡方以及显著性分析每条相互作用网络对患病程度的影响程度。

#### (1) 多元 Logistic 模型

设因变量有  $k$  个水平,可以对其中的  $k-1$  个水平各做一个回归方程,每个水平的因变量的概率取值范围为  $(0, 1)$ 。当自变量为连续变量时,就可以用 Logistic 方法对因变量的概率值建立回归方程。

对于因变量共有  $k$  个水平,以第  $k$  个水平为引用类别,则需要做  $k-1$  次二元 Logistic 回归,其中对于第  $j$  个水平的 Logistic 回归模型为;

$$\ln\left(\frac{p_j}{1-p_j}\right) = (\beta_0 + \sum_{i=1}^n \beta_{ji} x_{ji})$$

## (2) 拟合检验

Pearson 卡方统计量是用于检验多维表中观测频数与预测频数之间的差异。

定义为  $X^2 = \sum_{\text{所有单元格}} \frac{(\text{观测频数} - \text{预测频数})^2}{\text{预测频数}}$  36, Pearson 卡方统计量值越大, 说明拟合效果越不好。其次卡方偏差也可以检测模型拟合度, 如果模型拟合度好, 则对数似然比的差值就小, 显著性水平值越大;

似然比检验可以检验出某一变量与因变量的相关性有多大, 其零假设为某一因变量从模型中剔除后系数没有变化, 进一步可以根据显著性判断: 如果显著性小于 0.05, 即拒绝零假设, 可以认为这一变量从模型中剔除后系数发生变化, 显著性越小, 某一变量剔除后, 系数变化的可能性就越大, 同样可以认为这一变量与因变量的相关性大。

### 5.2.2 关键通路模型求解

将每个已知患病程度的样本的 66 条相互作用网络作为自变量, 每条相互作用网络的变化程度作为对应的值, 已知患病程度的样本共计 146 个。假设 ‘0’ 代表患病等级为 I, ‘1’ 代表患病等级为 II, ‘2’ 代表患病等级为 III, ‘3’ 代表患病等级为 IV (由开头假设 I A 和 I B 统称为 I, II A 和 II B 统称为 II, III A 和 III B 统称为 III, IV A 和 IV B 统称为 IV)。使 ‘0’ ‘1’, ‘2’, ‘3’ 为因变量, 将自变量和因变量录入 SPSS 中, 如图 4 所示:

	网络1	网络2	网络65	网络66	患病等级
1	.0000000000000000	.1666666666666667	.0	0	0
2	.0000000000000000	.0000000000000000	.0	0	0
3	.3333333333333333	.0555555555555556	.0	0	0
4	.0000000000000000	.0000000000000000	.0	0	0
5	.0000000000000000	.0000000000000000	.0	0	0
6	.0000000000000000	.0000000000000000	.0	0	0
7	.0000000000000000	.0000000000000000	.0	0	0
8	.4444444444444444	.4444444444444444	.0	0	0
9	.0000000000000000	.1666666666666667	.0	0	0
10	.0000000000000000	.2777777777777778	.0	0	0
11	.0000000000000000	.0833333333333333	.0	0	0
12	.0000000000000000	.0833333333333333	.0	0	0
13	.3333333333333333	.1388888888888889	.0	1	0
14	.0000000000000000	.3888888888888889	.0	0	0
15	.0000000000000000	.4444444444444444	.0	0	0
16	.0000000000000000	.7222222222222222	.0	0	0
17	.0000000000000000	.1666666666666667	.0	0	0
18	.0000000000000000	.0555555555555556	.0	0	0
19	.0000000000000000	.1944444444444445	.0	0	0
20	.7777777777777778	.8888888888888888	.0	1	0
21	.0000000000000000	.1944444444444445	.0	0	0

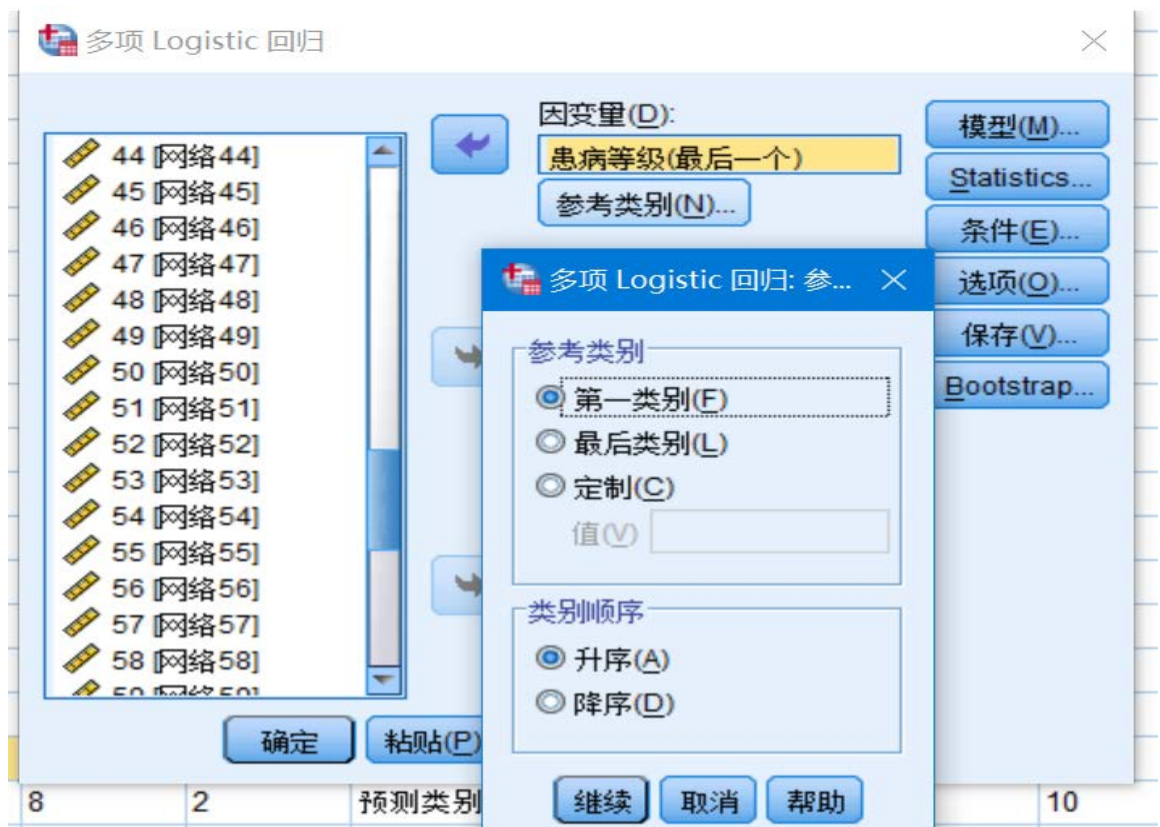


图 4

回归分析可以得到模型拟合度、伪 R 方和对数似然检验  
模型拟合信息：

模型拟合信息				
模型	模型拟合条件	似然比检验		
	-2 对数似然	卡方	自由度	显著性
仅有截距	388.703			
与前面连写	32.701	356.002	198	.000

图 5 模型拟合信息

从图 5 可以看出最终模型和模型中仅有截距项时的似然比检验结果，  
可以看到显著性小于 0.01，说明最终模型要优于仅有截距模型。

拟合优度：

拟合优度			
	卡方	自由度	显著性
皮尔逊	18.763	216	1.000
偏差	27.803	216	1.000

图 6 拟合优度

从图 6 可以看到 Pearson 统计量和偏差统计量，从显著性都大于 0.05，说明  
不能拒绝零假设，其中零假设为模型能很好拟合数据。

似然比检验(部分数据):

似然比检验				
效应	模型拟合条件	似然比检验		
	简化模型的 -2 对数似然	卡方	自由度	显著性
截距	76.375 <sup>a</sup>	43.675	3	.000
网络1	85.112 <sup>a</sup>	52.412	3	.000
网络2	18.928 <sup>b</sup>	.	3	.
网络3	54.304 <sup>a</sup>	21.603	3	.000
网络4	4.901 <sup>b</sup>	.	3	.
网络5	5164.976 <sup>a</sup>	5132.276	3	.000
网络6	4.904 <sup>b</sup>	.	3	.
网络7	4.902 <sup>b</sup>	.	3	.
网络8	55.733 <sup>a</sup>	23.032	3	.000
网络9	4.901 <sup>b</sup>	.	3	.
网络10	85.460 <sup>a</sup>	52.759	3	.000
网络11	454.670 <sup>a</sup>	421.969	3	.000
网络12	5.041 <sup>b</sup>	.	3	.
网络13	403.720 <sup>a</sup>	371.020	3	.000
网络14	82.776 <sup>a</sup>	50.075	3	.000
网络15	965.332 <sup>a</sup>	932.631	3	.000
网络16	58.886 <sup>a</sup>	26.186	3	.000
网络17	4.907 <sup>b</sup>	.	3	.
网络18	4.903 <sup>b</sup>	.	3	.

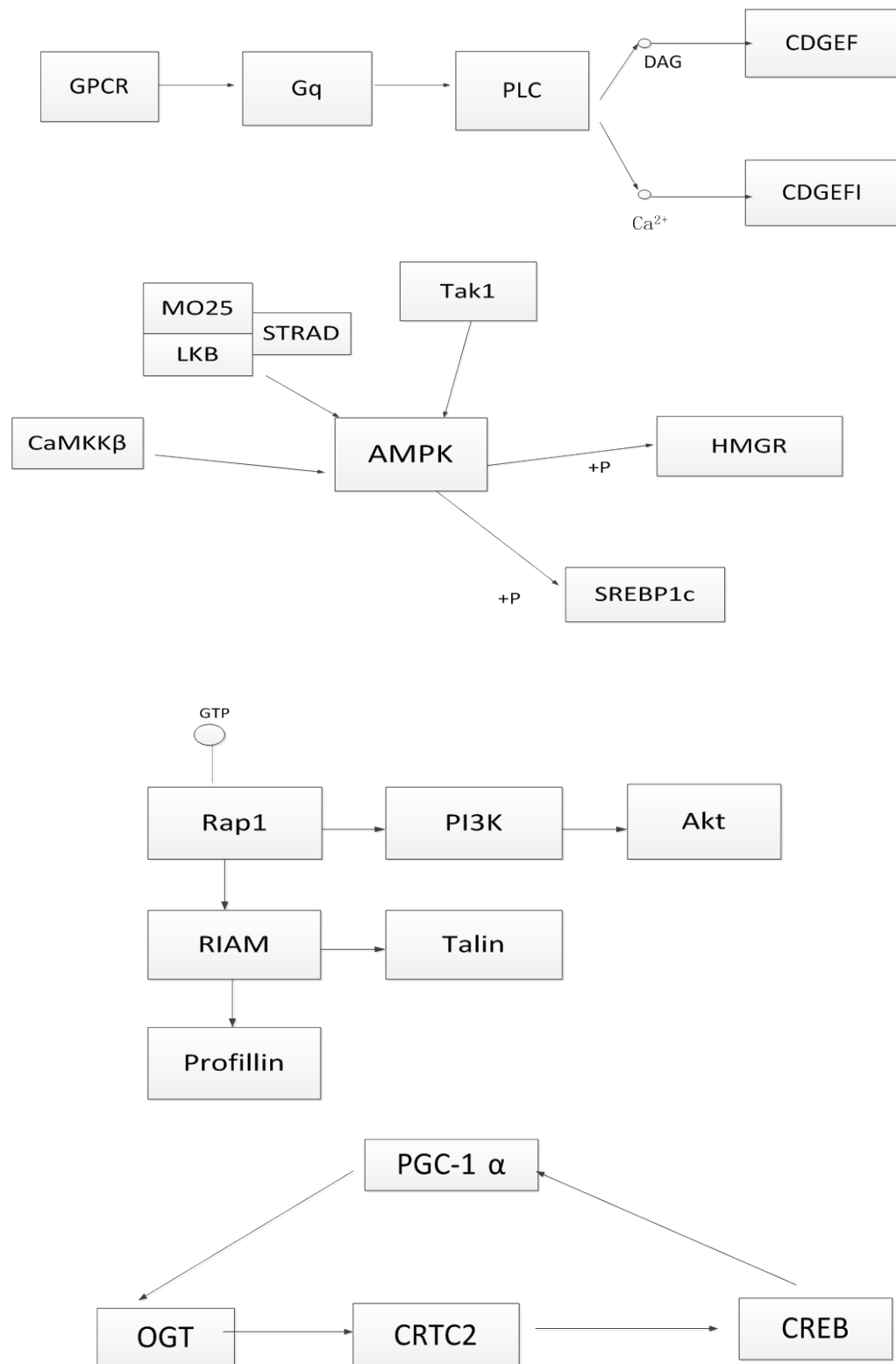
没有显著性的网络是由于最大似然值不存在或是拟似完全分隔,对此先将显著性小于 0.05 (对系数的影响是显著的) 的网络提取出来并按照卡方值的大小降序排列, 如表 9 所示:

	A	B	C	D	E
1	通路	对数似然	卡方	自由度	显著性
2	网络6	5164.976a	5132.276	3	0
3	网络47	1535.051a	1502.351	3	0
4	网络35	1420.668a	1387.968	3	1.2E-300
5	网络23	1135.990c	1103.29	3	7E-239
6	网络16	965.332a	932.6311	3	7.4E-202
7	网络37	690.220a	657.5191	3	3.4E-142
8	网络12	454.670a	421.9694	3	3.86E-91
9	网络14	403.720a	371.0198	3	4.19E-80
10	网络38	359.112a	326.4112	3	1.91E-70
11	网络53	305.622a	272.9219	3	7.2E-59
12	网络21	154.704c	122.0036	3	2.86E-26
13	网络27	132.590a	99.88989	3	1.64E-21
14	网络42	131.312a	98.61171	3	3.09E-21
15	网络36	98.742a	66.04096	3	3E-14
16	网络26	89.341a	56.64011	3	3.07E-12
17	网络11	85.460a	52.75921	3	2.06E-11
18	网络2	85.112a	52.41152	3	2.45E-11
19	网络15	82.776a	50.07505	3	7.7E-11
20	网络44	73.912a	41.21167	3	5.9E-09
21	网络50	73.074a	40.3732	3	8.88E-09
22	网络51	70.868a	38.16734	3	2.61E-08
23	网络54	63.730a	31.02969	3	8.38E-07
24	网络57	59.586a	26.88591	3	6.22E-06
25	网络17	58.886a	26.18591	3	8.78E-06

表 9

卡方值越大, 显著性就越小, 及更要拒绝零假设 (零假设为某因素变量从模型中剔除后系数不发生变化) 表明此网络相关性就越强。鉴于此, 我们共筛选出了 30 条网络作为代谢综合征的关键通路。表 9 标红部分为 30 条关键通路中的一部分。

图 7 为其中的 5 条关键通路图 (限于篇幅, 只展示部分相互作用网络):



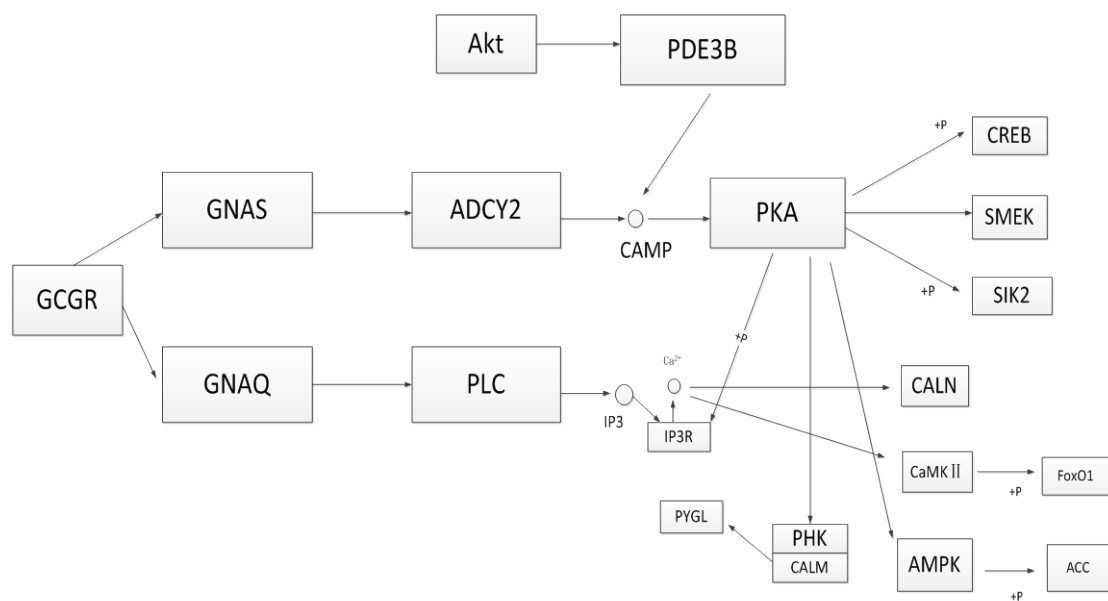


图 7 关键通路

### 5.2.3 关键通路的模型检验

为了检验所筛选的这 30 条关键通路，使用剔除后的 35 条相互作用网络作为变量再次使用多元 logistic 回归分析以检验模型得到似然比检验结果，如图 8 所示

似然比检验				
效应	模型拟合条件	似然比检验		
	简化模型的 -2 对数似然	卡方	自由度	显著性
截距	191.828 <sup>a</sup>	1.765	3	.623
网络3	191.167 <sup>a</sup>	1.103	3	.776
网络5	193.959 <sup>a</sup>	3.895	3	.273
网络7	221.727 <sup>a</sup>	31.663	3	.000
网络8	196.244 <sup>a</sup>	6.180	3	.103
网络10	190.865 <sup>a</sup>	.801	3	.849
网络13	190.651 <sup>a</sup>	.587	3	.899
网络18	197.961 <sup>a</sup>	7.897	3	.048
网络19	214.880 <sup>a</sup>	24.817	3	.000
网络20	202.180 <sup>a</sup>	12.117	3	.007
网络22	193.379 <sup>a</sup>	3.316	3	.345
网络24	191.372 <sup>a</sup>	1.309	3	.727
网络25	190.401 <sup>a</sup>	.337	3	.953
网络29	192.522 <sup>a</sup>	2.458	3	.483
网络30	194.333 <sup>a</sup>	4.270	3	.234
网络31	194.899 <sup>a</sup>	4.835	3	.184
网络32	197.475 <sup>a</sup>	7.411	3	.060
网络34	197.129 <sup>a</sup>	7.065	3	.070
网络40	207.349 <sup>a</sup>	17.285	3	.001

图 8 似然比检验结果

从图 8 可以看出剩余的相互作用网络的显著性绝大部分都大于 0.05（全部数据见附录），即要接受零假设（零假设为某因素变量从模型中剔除后系数不发生变化）。除此之外我们也可以得到他们的分类预测表：

观测值	预测值				正确百分比
	0	1	2	3	
0	44	1	2	5	84.6%
1	15	17	2	0	50.0%
2	10	0	16	2	57.1%
3	13	0	1	18	56.3%
总体百分比	56.2%	12.3%	14.4%	17.1%	65.1%

从分类表中也可以看出它的正确百分比只有 65.1%。

由以上我们认为这 30 条相互作用网络为关键通路。

### 5.3 问题三模型建立与求解

第一步：根据第一问 SAM 法获取的差异表达基因分别在两组别进行最后一次筛选，计算其假发现率，根据假发现率随阈值的变化情况选择合适的阈值。筛选出表达差异基因及其表达量 `gen_1`、`gen_2`。

第二步：将所选的 `gen_1`、`gen_2` 作为神经网络输入因子，分别对两组别样本建立基因表达量与患病严重程度关系的 BP 神经网络并将对应已知患病严重程度的部分数据带入进行训练。为增加训练样本容量，采取有放回的抽取。

第三步：将未带入训练的数据随机抽取 100 组（有放回的抽取）对所构建的 BP 神经网络进行测试。

第四步：将未知患病严重程度的样本相关数据代入训练好的 BP 神经网络中进行患病严重程度识别。

第五步：由第二问所得的关键通路作为第三个 BP 神经网络间的输入因子，患病严重程度作为输出层，代入已知的部分数据进行训练，剩余部分随机抽取 100 组作为测试。

第六步：将未知样本的相关数据代入 BP 神经网络进行预测，得到由通路变化程度预测的患病严重程度。

流程图如图 9 所示



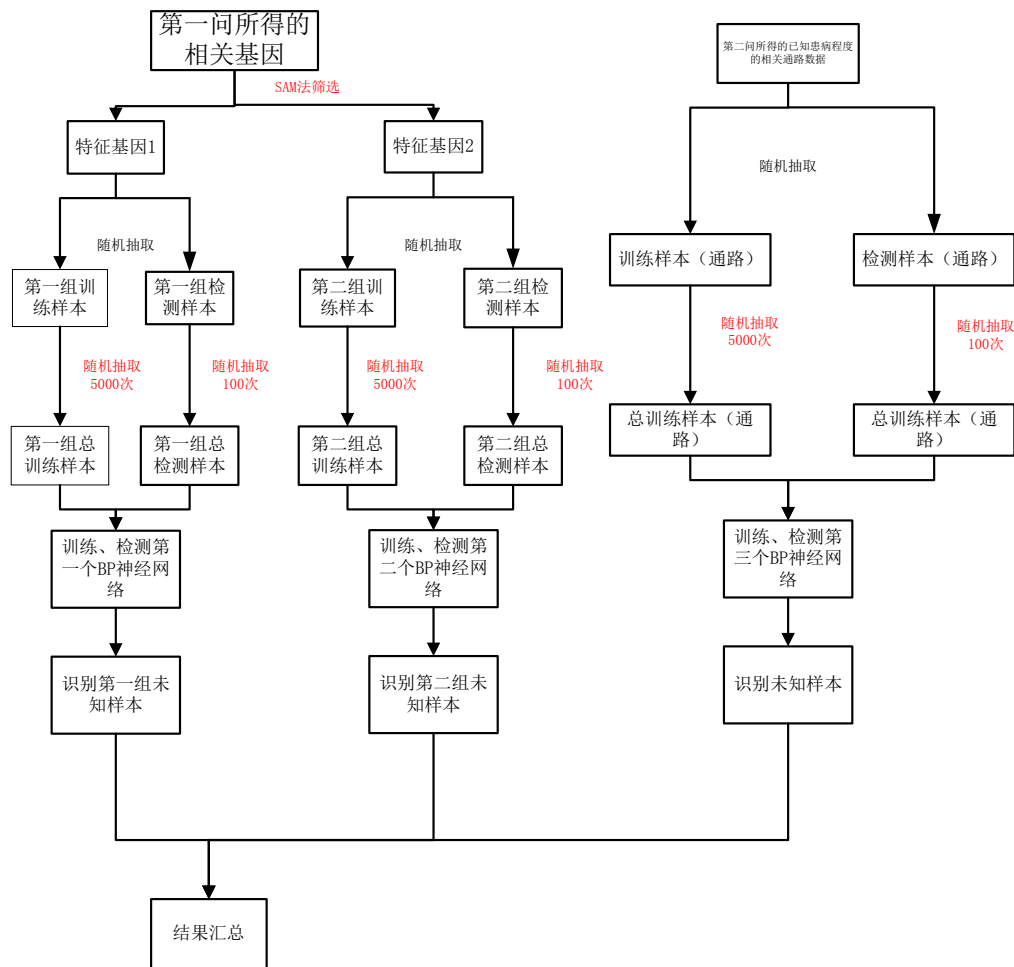


图 9 第三问流程图

### 5.3.1 BP 神经网络输入因子的选取

根据第一问 SAM 法获取的差异表达基因分别在两组别进行最后一次筛选，计算其假发现率，根据假发现率随阈值的变化情况选择合适的阈值。筛选出特征基因 gen\_1、gen\_2。（如表 10、11 所示）

AP1G1 164	LOC200030 200030	SLC25A10 1468	KAL1 3730
C1RL 51279	MAFK 7975	TNFSF13 8741	SAP30BP 29115
CXCL14 9547	MAMDC2 256691	UBC 7316	
FAM65B 9750	PDP2 57546	ZMYM3 9203	
HTRA2 27429	RNPC3 55599	ZNF510 22869	

表 10 第一批样本神经网络输入因子选取结果

ANK1 286	EMP2 2013	PRRG1 5638	TBRG1 84897
ARMCX1 51309	GMPPB 29925	RASL11A 387496	TRIM6 117854
C12orf39 80763	KIAA1279 26128	RHBDL2 54933	ZNF324B 388569
C9orf153 389766	ODF2L 57489	SDC2 6383	ZNF727 442319

表 11 第二批样本神经网络输入因子选取结果

以选出的特征基因 **gen\_1**（表 10）、**gen\_2**（表 11）分别作为第一批样本神经网络和第二批样本神经网络的输入因子。

根据第二问所得的关键通路作为第三个 BP 神经网络间的输入因子，患病严重程度作为输出层。

### 5.3.2 各 BP 神经网络隐含层节点个数确定

对于样本患病严重程度识别的三个神经网络，设定它们的隐含层层数为 3，依据经验公式：

$$M = \sqrt{n + m} + a^{[1]}$$

（M：隐含层节点个数，n:输入层神经元个数，m: 输出层神经元个数，a:0~10 之间的常数）

第一批样本 BP 神经网络隐含层节点个数确定如下：

隐含层	节点个数
第一层	15
第二层	12
第三层	5

表 12 第一批样本 BP 神经网络隐含层节点个数

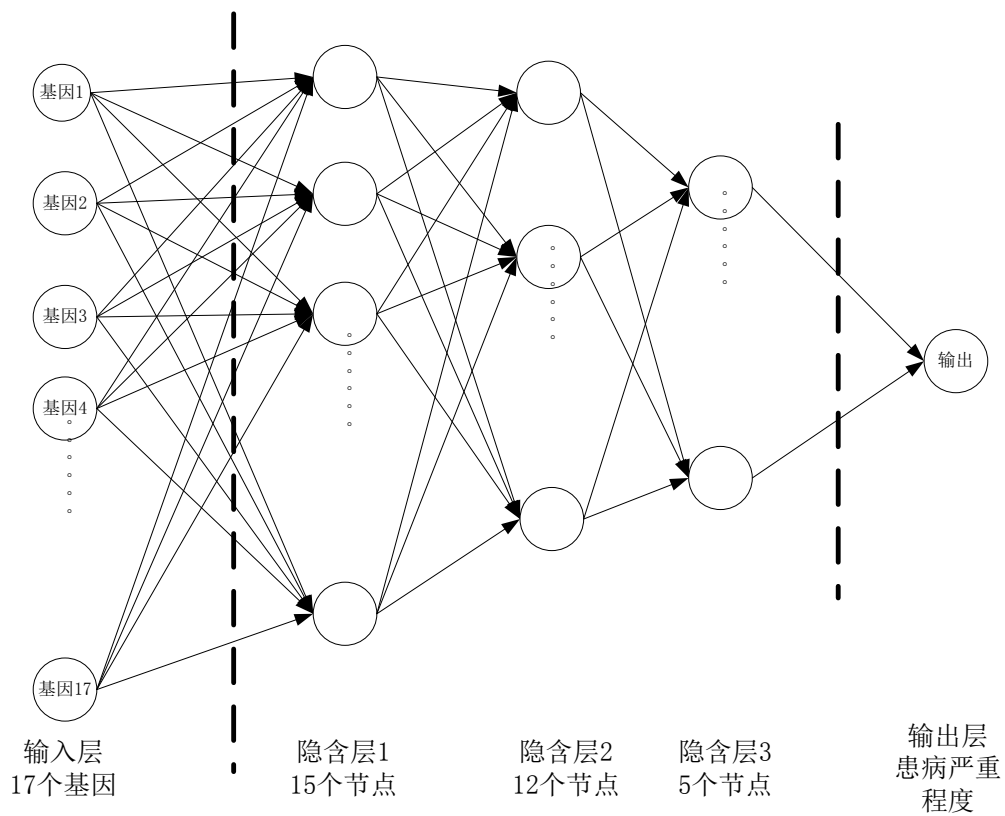


图 10 第一批样本表达量与患病严重程度之间的 BP 神经网络结构图

第二批样本 BP 神经网络隐含层节点个数确定如下：

隐含层	节点个数
第一层	15
第二层	12
第三层	5

表 13 第二批样本 BP 神经网络隐含层节点个数

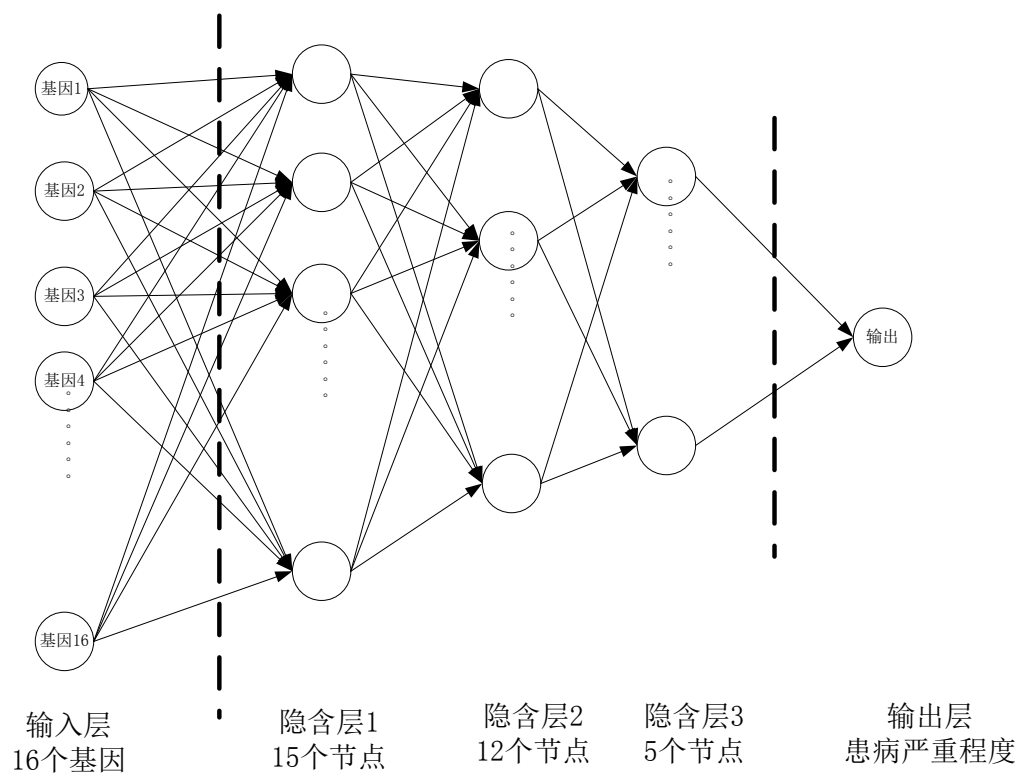


图 11 第二批样本表达量与患病严重程度之间的 BP 神经网络结构图

通路变化情况与患病严重程度之间 BP 神经网络隐含层节点个数确定如下：

隐含层	节点个数
第一层	17
第二层	10
第三层	8

表 14 通路变化情况与患病严重程度之间 BP 神经网络隐含层节点个数

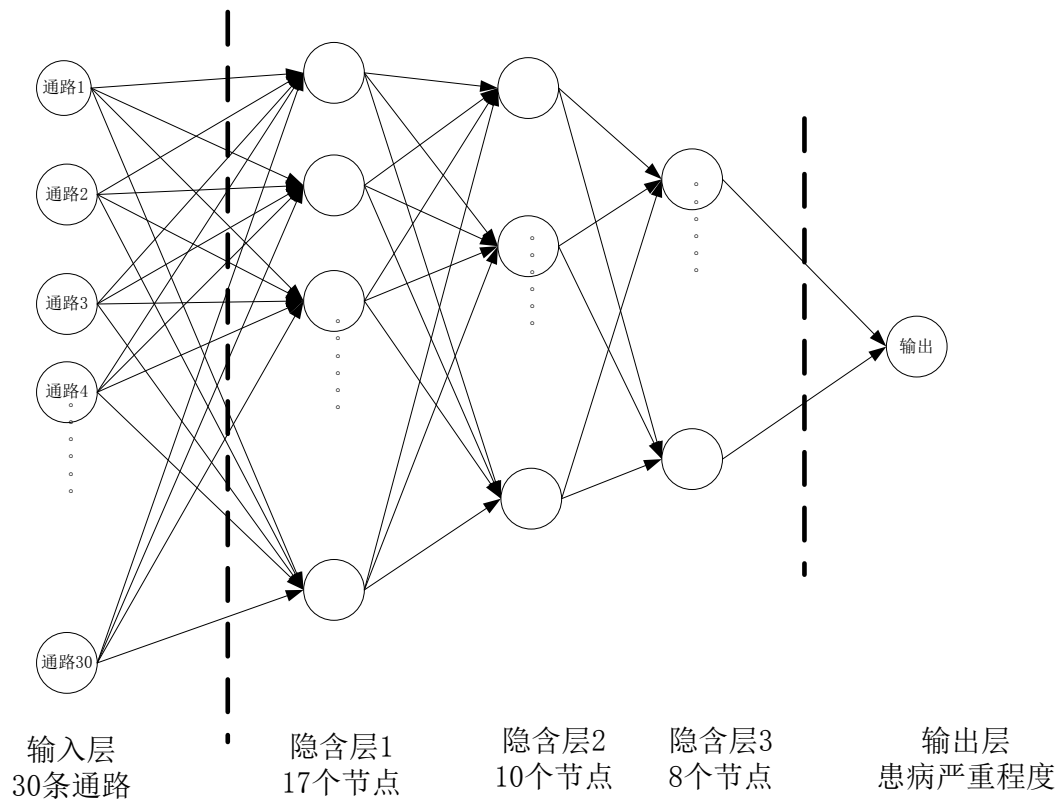


图 12 通路变化情况与患病严重程度之间的 BP 神经网络结构图

### 5.3.3 模型求解

第一批样本数据 BP 神经网络的建立及训练

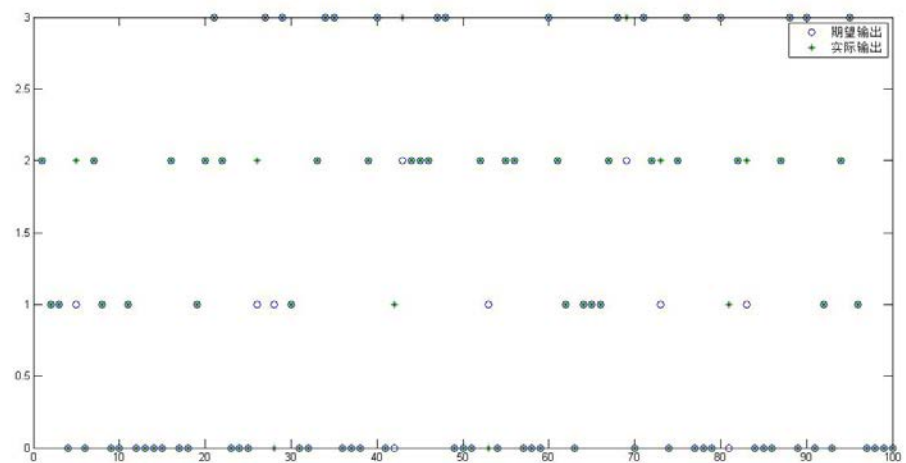


图 13 第一批样本 BP 神经网络测试结果

图 13 可以直观看出期望输出‘o’与实际输出‘+’基本吻合，仅有少部分识别错误，通过计算可得其识别准确率为 86%，所以可以认为第一批样本 BP 神经网络建立正确。

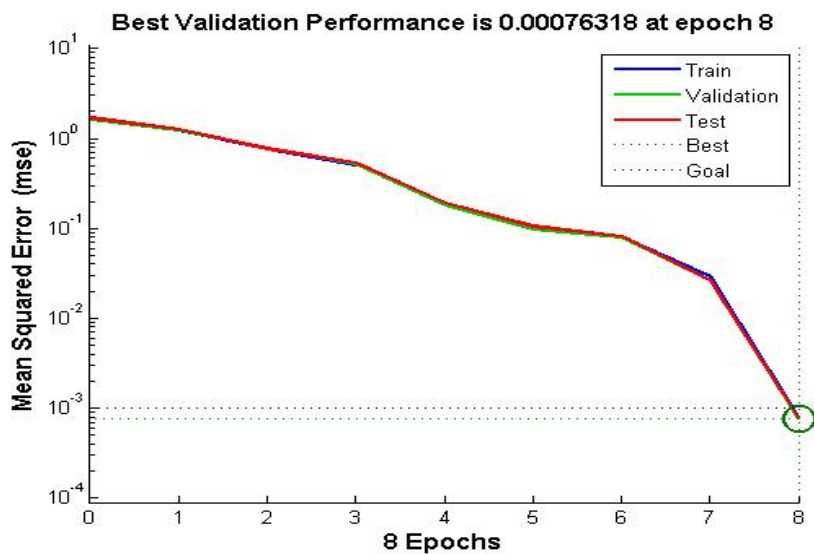


图 14 第一批样本 BP 神经网络训练过程

将第一批样本中待预测数据输入训练好的神经网络进行识别，最终结果在结果汇总展示（表 15）

第二批样本数据 BP 神经网络的建立及训练

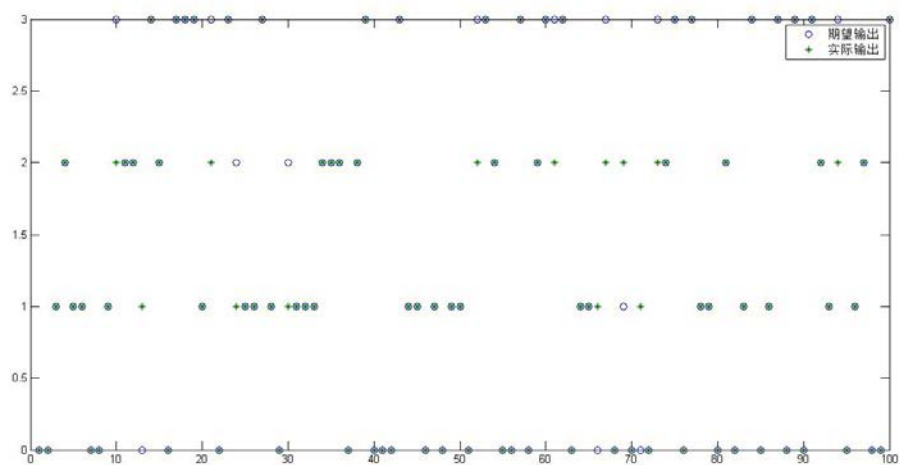


图 15 第二批样本 BP 神经网络测试结果

图 15 可以直观看出期望输出‘o’与实际输出‘+’基本吻合，仅有少部分识别错误，通过计算可得其识别准确率为 89%，所以可以认为第二批样本 BP 神经网络建立正确。

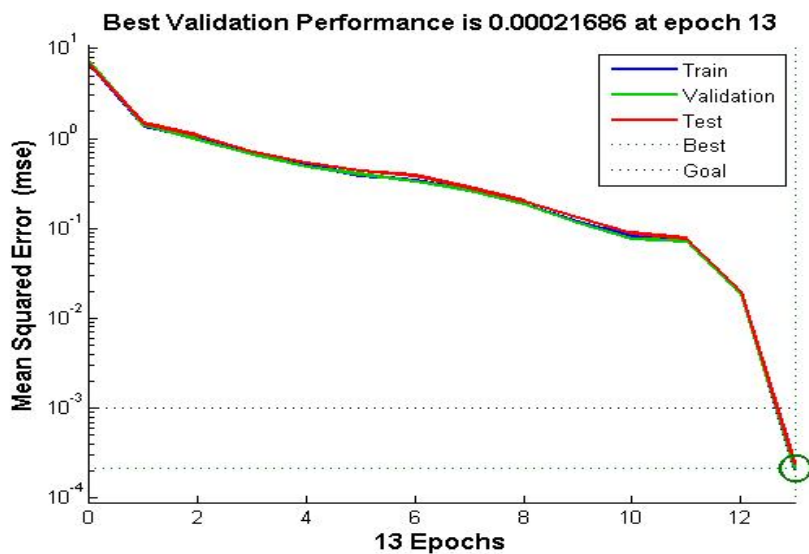


图 16 第二批样本神经网络训练过程

将第二批样本中待预测数据输入训练好的神经网络进行识别，最终结果在结果汇总展示。（表 15）

通路变化程度与患病严重程度之间的 BP 神经网络建立

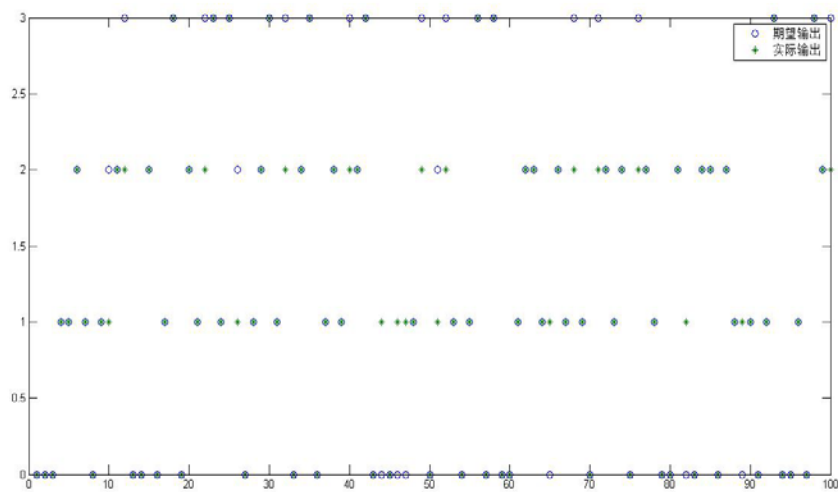


图 17 通路变化程度与患病严重程度之间的 BP 神经网络测试结果

图 17 可以直观看出来期望输出‘o’与实际输出‘\*’基本吻合，仅有少部分识别错误，通过计算可得其识别准确率为 80%，通路变化程度与患病严重程度之间的 BP 神经网络建立正确。

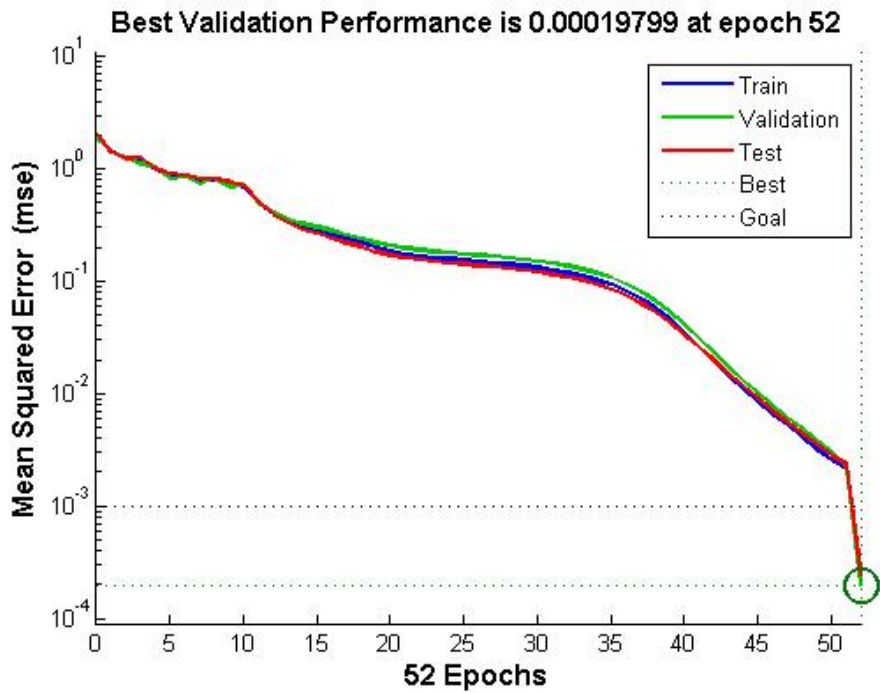


图 18 通路变化程度与患病严重程度之间的 BP 神经网络训练过程

将第一批次和第二批次所有未知患病程度样本输入训练好的神经网络进行识别，最终结果在汇总中展示（表 15）

由前两个 BP 神经网络可以得到未知患病严重程度的样本的现状，由第三个 BP 神经网络可以得到未知样本未来患病风险。

通过以上步骤，我们的得到如表 15 所示结果

	A	B	C	D	E	F	G	H	I
	编号	患病严重程度(RNA)	患病风险(通路情况)	编号	患病严重程度(RNA)	患病风险(通路情况)	编号	患病严重程度(RNA)	患病风险(通路情况)
1	1	3	2	29	1	1	57	0	3
2	2	3	1	30	2	1	58	0	1
3	3	1	0	31	2	0	59	0	1
4	4	0	3	32	1	1	60	0	1
5	5	0	1	33	0	0	61	2	0
6	6	0	0	34	0	0	62	1	3
7	7	1	0	35	2	1	63	1	1
8	8	0	1	36	0	0	64	1	0
9	9	0	0	37	2	0	65	2	1
10	10	0	1	38	1	1	66	0	2
11	11	2	0	39	2	1	67	1	3
12	12	2	1	40	0	0	68	0	1
13	13	2	0	41	1	1	69	2	0
14	14	0	1	42	2	1	70	3	1
15	15	1	1	43	2	0	71	1	0
16	16	1	3	44	2	0	72	1	2
17	17	1	2	45	2	1	121	0	1
18	18	2	0	46	2	1	122	2	0
19	19	1	1	47	0	0	123	3	1
20	20	2	0	48	1	0	124	0	1
21	21	1	1	49	0	2	125	1	2
22	22	1	0	50	1	1	126	3	2
23	23	0	2	51	0	1	127	0	1
24	24	0	0	52	0	1	128	3	3
25	25	0	3	53	1	1	129	2	1
26	26	2	1	54	1	1	130	2	3
27	27	2	1	55	1	1	131	0	0
28	28	0	1	56	0	1			

表 15 预测所得汇总结果

根据预测的情况分析关键通路变化情况找出样本患代谢综合征风险的原因如表 16 所示（限于篇幅，只展示部分）



	A	B	C	D
1	编号	患病严重程度(RNA)	患病风险(关键通路情)	原因
2	1	3	2	关键通路2、4、6、11、12、14、15、23、27、33、35、38、44、50、51、53、56、57发生变化
3	2	3	1	关键通路2、11、14、15、23、33、36、38、39、44、47、50、53、54、56发生变化，变化程度不高
4	3	1	0	关键通路基本无变化
5	4	0	3	关键通路2、15、23、38、47、56发生变化
6	5	0	1	关键通路2、11、12、14、15、23、26、27、33、36、39、44、50、51、53、56、57发生变化，变化程度不高
7	6	0	0	关键通路基本无变化
8	7	1	0	关键通路基本无变化
9	8	0	1	关键通路56、57发生变化，变化程度不高
10	9	0	0	关键通路基本无变化
11	10	0	1	关键通路23发生变化
12	11	2	0	关键通路基本无变化
13	12	2	1	关键通路56发生变化
14	13	2	0	关键通路基本无变化
15	14	0	1	关键通路2、4、11、23、33、36、39、44、50、56发生变化，变化程度不高
16	15	1	1	关键通路基本无变化
17	16	1	3	关键通路23发生变化，且变化程度高
18	17	1	2	关键通路11、16、33、36、39、44、50、56发生变化
19	18	2	0	关键通路基本无变化
20	19	1	1	关键通路基本无变化
21	20	2	0	关键通路基本无变化
22	21	1	1	关键通路基本无变化
23	22	1	0	关键通路基本无变化
24	23	0	2	关键通路2、11、23、36、39、44、50发生变化
25	24	0	0	关键通路基本无变化
26	25	0	3	关键通路68、56发生变化
27	26	2	1	关键通路2、11、47、54、

表 16 样本患病风险原因分析

基于前文假设其中 0 表示 stage I

1 表示 stage II

3 表示 stage III

4 表示 stage IV

由上表可以看出有些样本根据 DNA 表达量预测的患病严重程度与根据通路变化情况预测的患病严重程度不同，其中由 DNA 表达量预测的患病严重程度为样本的现状，根据通路变化情况预测样本未来患病风险。

个性化治疗方案：（限于篇幅，只展示部分）

编号	DNA 表达量预测结果	通路变化情况风险预测	个性化方案
1	3	2	注意日常生活习惯
2	3	1	注意日常生活习惯
3	1	0	注意日常生活习惯
4	0	3	虽然现状为未患病，但患病风险大，应接受基因治疗。

5	0	1	虽然现状为未患病，但患病风险大，应接受基因治疗。
6	0	0	保持现状。
7	1	0	注意日常生活习惯
8	0	1	虽然现状为未患病，但患病风险大，应接受基因治疗。
9	0	0	保持现状。
10	0	1	虽然现状为未患病，但患病风险大，应接受基因治疗。
11	2	3	病情有很大风险恶化，应接受基因治疗。
12	2	1	注意日常生活习惯
13	2	3	病情有很大风险恶化，应接受基因治疗。
14	0	1	虽然现状为未患病，但患病风险大，应接受基因治疗。
15	1	1	注意日常生活习惯，并接受一定的治疗
16	1	3	病情有很大风险恶化，应接受基因治疗。
17	1	2	病情有很大风险恶，应接受基因治疗。化
18	2	0	注意日常生活习惯
19	1	1	注意日常生活习惯，并接受一定的治疗
20	2	0	注意日常生活习惯
21	1	1	注意日常生活习惯，并接受一定的治疗

22	1	0	注意日常生活习惯
23	0	2	虽然现状为未患病，但患病风险大，应接受基因治疗。
24	0	0	保持现状。
25	0	3	虽然现状为未患病，但患病风险大，应接受基因治疗。
26	2	1	注意日常生活习惯
27	2	1	注意日常生活习惯

表 17 相应的治疗建议

## 六、模型评价

### 6.1 模型优点

1. 从基因表达量文件中筛选出与代谢综合征有关的相关基因，采用了 SAM 法，根据每次 SAM 法筛选后的 3 个参数（统计表达量  $d_0$ 、全排序法所得的统计表达量  $d$ 、迭代次数），进行假阳性检验，分析假发现率随阈值的变化趋势，选取合适的阈值。同时保证较多的相关基因被筛选与低假发现率。
2. 在 3 个 BP 神经网络构建过程中再次对相关基因进行 SAM 法筛得到特征基因，选取第二问所得的关键通路，减少无关变量对网络的影响程度。
3. 对训练数据进行归一化处理去除数据量纲，使数据大小均在  $[0, 1]$  范围内，加快了训练速度，同时在训练样本中分别加入了强度为 0.1、0.05、0.001 的噪声，有效避免了神经网络过拟合出现，增强 BP 神经网络的泛化能力。

### 6.2 模型缺点

- 1.通过 KEGG 检索相关生物分子作用网络忽略了尚未发现的有关作用网络。只模糊评价网络中各基因的重要性,未深入分析相关基因变异对网络的具体影响,在识别方面存在一定的误差。
- 2.在 SAM 法筛选过程占用大量时间(每次耗时 30 秒左右)。
- 3.由于训练样本太少的局限性,对训练样本随机抽样,不能完全代表总统样本,在识别方面存在一定的误差。
- 4.未考虑相关通路间的作用,导致样本识别方面存在一定误差。

## 参考文献

- [1] 闻新等,应用 MATLAB 实现神经网络,国防工业出版社,2015: 95 -160.
- [2] 刘丹红等.临床疾病严重度评价标准.中国医学统计.2002.9( 2 ): 78-80.
- [3] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response [J]. Proceeding of the National Academy of Sciences , 2001, 98(9): 5116-5121
- [4] 王小燕. 两样本问题中差异表达基因的筛选研究.中南大学硕士学位论文.2012
- [5] 康海静等.代谢综合征的研究进展. 医学综述. 2007;13:68-70.
- [6] 许晓君等.广东省 15 岁以上人群代谢综合征流行特征分析[期刊论文]-华南预防医学.2009(05).
- [7] 互动百科 <http://www.baik.com/wiki/代谢综合征>

## 附录

注意事项:

1. 此题程序均在 MATLAB 2013a, win7 64 位环境下运行;
2. 相关基因筛选过程在 Excel 中进行;
3. 相互作用网络中各基因的权值已随程序给出

## 附录一 SAM 法筛选相关基因函数

```
function [tezheng, d, d0] = my_sam(data, num1, num2, num3)

% 作用：从基因表达数据中筛选出相关基因

% 输入数据：data 基因表达数据、num1 类别一的个数、num2 类别二的个数、num3 迭代次数

%输出数据：d0 未进行 permutation 所计算的统计表达量，d 为其中的 dp

%  $a = (1/num1 + 1/num2) / (num1 + num2 + 2)$ 

a = (1/num1 + 1/num2) / (num1 + num2 + 2);

num = num1 + 1;

[~, N2] = size(data);

data1 = data(:, 1:num1);

data2 = data(:, num:N2);

data1 = data1';

data2 = data2';

u1 = mean(data1);

u2 = mean(data2);

u1 = u1';

u2 = u2';

K1 = u1 - u2;

s1 = var(data1);

s2 = var(data2);

s1 = s1';

s2 = s2';

s1 = s1 * (num1 - 1);

s2 = s2 * (num2 - 1);

S1 = (a * (s1 + s2)) . ^ (1/2);

c = [];

%% 寻找最大 S

S_max = 0;

for i = 1:2000;
```

```

new = randperm(N2);

data = data(:,new);

data1 = data(:,1:num1);

data2 = data(:,num:N2);

data1 = data1';

data2 = data2';

s1 = var(data1);    %计算方差，根据方差推出偏方差

s2 = var(data2);

s1 = s1*(num1-1);

s2 = s2*(num2-1);

s1 = s1';

s2 = s2';

zhangshi = (a*(s1+s2)).^(1/2);

S_max = max(S_max, zhangshi);

end

d0 = K1./(S1+S_max);          %计算原始概率统计量

%% 打乱顺序重新计算

d=[];

for i = 1:num3;

    new = randperm(N2);

    data = data(:,new);

    data1 = data(:,1:num1);

    data2 = data(:,num:N2);

    data1 = data1';

    data2 = data2';

    K = (mean(data1) - mean(data2));

    K = K';

    s1 = var(data1);

    s2 = var(data2);

```

```

s1 = s1*(num1-1);
s2 = s2*(num2-1);
S = (a*(s1+s2)).^(1/2);
S = S';
d_z = K./(S+S_max);
data1 = data1';
data2 = data2';
d = [d, d_z];
end
d_mean = d';
d_mean = mean(d_mean);
d_mean = d_mean';
tezheng = abs(d0(:) - d_mean(:));

```

## 附录二 假阳性检验函数

```

function rate = fdr_test(d0,d,tezheng,num3)
% 目标:输入 SAM 法所得的统计表达量与迭代次数, 得到假发现率随阈值变化趋势
%输入: d0 未进行 permutation 所计算的统计表达量, d 为其中的 dp, num3 为 SAM 迭代次数
%输出: rate 阈值与对应的假发现率
rate = [];
c = [];
for yuzhi = 0.1:0.1:2
    d_t = d0(find(tezheng>yuzhi));
    d_p = d(find(tezheng>yuzhi),:);
    zhongshu = size(d_t,1);
    for i = 1:num3
        d_q= d_p(:,i);

```

```

        n = sum(abs(d_q)>abs(d_t));
        c = [c,n];
    end
c = sum(c)/num3;
FDR = c/zhongshu;
rate = [rate,FDR];
end
yuzhi = 0.1:0.1:2;
plot(yuzhi,rate,'*-');
place = find(rate<0.005);
rate = [yuzhi;rate];
rate = rate';
% place = place(1);
title(' I 与 I I 间假阳性检验');
xlabel(' 阈值');
ylabel(' 假发现率');
text(yuzhi(place),rate(place),[' (' ,num2str(yuzhi(place)),', ',',num2str(
rate(place)),', ')] , 'color','b');

```

### 附录三 计算各样本相互作用网络变化程度函数

```

function data_output = pathpinjia(path,var_inf)
% 目的:根据样本变异信息, 计算每个样本的相互作用网络的变化程度
%输入:path 相互作用网络中每个基因及其对应的权值, var_inf 样本变异信息
%输出 data_output
[~,N1] = size(path);%大小
data_output = zeros(229,N1);%大小
for i = 1:229
    data = [];
    place = find(var_inf.num==i);%第 num 个人

```



```

str = var_inf.nam(place);%中间变量
[N2,~] = size(str);
for j = 1:N2
    str1 = str(j);
    for k = 1:N1
        place1 = find(strcmp(path(k).nam, str1));%找位置
        if ~isempty(place1)%判空
            w(k) = sum(path(k).weight(place1));%赋值
        else
            w(k) = 0;
        end
        zhong(k) = sum(path(k).weight);
    end
    data = [data;w];%累加
end
data = sum(data);
data_output(i,:) = data./zhong;
clear data;
end

```

#### 附录四 随机数列产生函数

```

function a = menterkal(train_data)
a = [];
[~,n] = size(train_data);%计算矩阵大小
N = ceil(5000/n);%取整
for i = 1:N;
    x = randperm(n);%产生随机序列
    a = [a x];
end

```

## 附录五 相关数据检索函数

```
function data_output = my_look(name, data_name, data1)
%name 基因名 data1 匹配数据
%str 暂存基因名
[N, ~] = size(name);
place = [];
for i = 1:N
    str = name(i);
    place1 = find(strcmp(data_name, str));
    place = [place; place1];
end
data_output = data1(place, :);
```

## 附录六 第一批样本 BP 神经网络建立、训练、测试

```
%% 分配数据
%数据说明：第一行为样本编号，中间行为相关输入数据，最后一行为样本对应
患病严重程度
% 目的：将数据对半分为训练数据与测试数据
[x1, y1] = size(sample);
a = menterkal(sample);
n = ceil(y1/2);
test_in = sample(2:end-1, a(1:n));
test_true = sample(end, a(1:n));
sample1 = [test_in; test_true];
train_in = sample(2:end-1, a(n+1:2*n));
[x2, y2] = size(train_in);
noisy = rand(x2, y2)*0.1;
train_in = noisy+train_in; %加入噪声
train_out = sample(end, a(n+1:2*n));
```

```

%% 随机抽取实验数据
%目的：将训练样本抽至 5000 多个
sample2 = [train_in;train_out];
a = menterkal(sample2);
train_in = sample2(1:end-1,a);
train_in(end+1,:) = 1;%训练数据加入阈值
train_out = sample2(end,a);
%% BP 神经网络（关于相关基因 DNA 表达与患病严重程度的导出关系）的建立
net = newff(train_in,train_out,[15 12
5],{'logsig','logsig','logsig'},'trainlm');
net.trainParam.show = 10;
net.trainParam.lr = 0.01;
net.trainParam.epochs = 1000;
net.trainParam.goal = 0.001;
%% BP 神经网络训练
[net,tr] = train(net,train_in,train_out);
test_in(end+1,:) = 1;%测试数据加入阈值
%% BP 神经网络测试
test_in = sample1(1:end-1,a);
test_in(end+1,:) = 1;%测试数据加入阈值
test_true = sample1(end,a)
test_out = sim(net,test_in);
x = 1:100;
test_out = my_sort(test_out);
y1 = test_true(1,a(1:100));%抽取 100 组检验
y2 = test_out(1,a(1:100));
plot(x,y1,'o',x,y2,'*');
legend('期望输出','实际输出');
t_rate = sum(test_out == test_true)/size(test_out,2) %识别正确率计算

```

## 附录七 第二批样本 BP 神经网络建立、训练、测试

```
%% 分配数据
%数据说明：第一行为样本编号，中间行为相关输入数据，最后一行为样本对应
患病严重程度
% 目的：将数据对半分为训练数据与测试数据
[x1,y1] = size(sample);
a = menterkal(sample);
n = ceil(y1/2);
test_in = sample(2:end-1,a(1:n));
test_true = sample(end,a(1:n));
sample1 = [test_in;test_true];
train_in = sample(2:end-1,a(n+1:2*n));
[x2,y2] = size(train_in);
noisy = rand(x2,y2)*0.01;
train_in = noisy+train_in; %加入噪声
train_out = sample(end,a(n+1:2*n));
%% 随机抽取实验数据
%目的：将训练样本抽至 5000 多个
sample2 = [train_in;train_out];
a = menterkal(sample2);
train_in = sample2(1:end-1,a);
train_in(end+1,:) = 1;%训练数据加入阈值
train_out = sample2(end,a);
%% BP 神经网络（关于相关基因 DNA 表达与患病严重程度的导出关系）的建立
net = newff(train_in,train_out,[15 12
5],{'logsig','logsig','logsig'},'trainlm');
net.trainParam.show = 10;
net.trainParam.lr = 0.05;
net.trainParam.epochs = 1000;
```

```

net.trainParam.goal = 0.001;
%% BP 神经网络训练
[net, tr] = train(net, train_in, train_out);
test_in(end+1, :) = 1;%测试数据加入阈值
%% BP 神经网络测试
test_in = sample1(1:end-1, a);
test_in(end+1, :) = 1;%测试数据加入阈值
test_true = sample1(end, a)
test_out = sim(net, test_in);
x = 1:100;
test_out = my_sort(test_out);
y1 = test_true(1, a(1:100));%抽取 100 组检验
y2 = test_out(1, a(1:100));
plot(x, y1, 'o', x, y2, '*');
legend('期望输出', '实际输出');
t_rate = sum(test_out == test_true)/size(test_out, 2) %识别正确率计算

```

#### 附录八 通路变化情况与患病严重程度之间 BP 神经网络建立、训练、测试

```

%% 分配数据
%数据说明：第一行为样本编号，中间行为相关输入数据，最后一行为样本对应
患病严重程度
% 目的：将数据对半分为训练数据与测试数据
[x1, y1] = size(sample);
a = menterkal(sample);
n = ceil(y1/2);
test_in = sample(2:end-1, a(1:n));
test_true = sample(end, a(1:n));
sample1 = [test_in; test_true];
train_in = sample(2:end-1, a(n+1:2*n));

```

```

[x2,y2] = size(train_in);
noisy = rand(x2,y2)*0.001;
train_in = noisy+train_in; %加入噪声
train_out = sample(end,a(n+1:2*n));
%% 随机抽取实验数据
%%目的：将训练样本抽至 5000 多个
sample2 = [train_in;train_out];
a = menterkal(sample2);
train_in = sample2(1:end-1,a);
train_in(end+1,:) = 1;%训练数据加入阈值
train_out = sample2(end,a);
%% BP 神经网络（关于相关基因 DNA 表达与患病严重程度的导出关系）的建立
net = newff(train_in,train_out,[17 10
8],{'logsig','logsig','logsig'},'trainlm');
net.trainParam.show = 10;
net.trainParam.lr = 0.01;
net.trainParam.epochs = 1000;
net.trainParam.goal = 0.001;
%% BP 神经网络训练
[net,tr] = train(net,train_in,train_out);
test_in(end+1,:) = 1;%测试数据加入阈值
%% BP 神经网络测试
test_in = sample1(1:end-1,a);
test_in(end+1,:) = 1;%测试数据加入阈值
test_true = sample1(end,a)
test_out = sim(net,test_in);
x = 1:100;
test_out = my_sort(test_out);
y1 = test_true(1,a(1:100));%抽取 100 组检验

```

```

y2 = test_out(1, a(1:100));
plot(x, y1, 'o', x, y2, '*');
legend('期望输出', '实际输出');
t_rate = sum(test_out == test_true)/size(test_out, 2) %识别正确率计算

```

## 附录九 第二批样本相关基因及假阳性检验结果

### I 与 II 间相关基因

###probe	差异性
ZNF222 7673	2.455342026
UBE2MP1 606551	2.422019698
RPL23P8 222901	2.268500509
LOC723972 723972	2.258015164
ANXA2P3 305	2.24321511
ANXA2P1 303	2.237775072
LOC649330 649330	2.21549304
ZNF439 90594	2.206286203
LOC442459 442459	2.180099063
CGRRF1 10668	2.166521575
RPL19P12 100129424	2.165256719
NACA2 342538	2.158712236
SPRED2 200734	2.122731914
NME2P1 283458	2.109001141
LOC407835 407835	2.098025093
PA2G4P4 647033	2.089261589
LOC401010 401010	2.088957072
H3F3C 440093	2.086142742
TCHHL1 126637	2.076924368
ATP5EP2 432369	2.075732728
ZNF582 147948	2.074550569
UBE2NL 389898	2.068199949
MORF4 10934	2.067869075
FAM102B 284611	2.061922798
ZSCAN12 9753	2.060132276
C2orf42 54980	2.056085703
HSPA4L 22824	2.055058496
CALHM2 51063	2.047144764
CYP20A1 57404	2.046395017

HSP90AB2P 391634	2.032575222
CCDC121 79635	2.031701789
LENG9 94059	2.031376325
ZNF740 283337	2.027125116
DHFRL1 200895	2.021716761
ZNF284 342909	2.020785566
ZSCAN2 54993	2.0205064
ZNF30 90075	2.018498634
SPACA5 389852	2.017405602
C14orf19 280655	2.017016506
PIPSL 266971	2.008949121
ZNF549 256051	2.007627951
ANXA2P2 304	2.006551721
NR3C2 4306	2.001821052
PTTG2 10744	1.998186277
PIP5K1P1 206426	1.997679296
MERTK 10461	1.996800651
ZNF2 7549	1.996065912
SUSD2 56241	1.994803533
PLXND1 23129	1.992391129
ZNF641 121274	1.992377088
C3orf18 51161	1.992311912
OR2A7 401427	1.990955399
ANP32C 23520	1.989099568
CSNK1A1L 122011	1.988434425
ZNF223 7766	1.988170138
ZNF708 7562	1.983863722
ANKRD40 91369	1.980246146
RASSF3 283349	1.977588209
CSDAP1 440359	1.975700701
SUMO1P3 474338	1.963993738
ZNF14 7561	1.961471269
ZIK1 284307	1.959651787
MGC16142 84849	1.95642891
SENP7 57337	1.955847868
ITGA9 3680	1.955286125
NACC1 112939	1.953696145
ANKRD57 65124	1.953244826
A2M 2	1.949178238
DLC1 10395	1.944269882
BTBD11 121551	1.936353383
ABI2 10152	1.933390267
PTTG3P 26255	1.929654764
HSP90B3P 343477	1.927757425



ZFP30 22835	1.926993314
ZNF792 126375	1.92428779
KLHL23 151230	1.922578555
GNG7 2788	1.921371705
CBX4 8535	1.909011928
C1QTNF7 114905	1.904272706
LMNA 4000	1.902914089
EXOC6 54536	1.902880685
SIDT2 51092	1.900495199
PRPS1L1 221823	1.89921996
TCF21 6943	1.897860496
RNF125 54941	1.89762893
ZNF419 79744	1.895279997
WBP11P1 441818	1.891096537
CARM1 10498	1.890468762
ZBTB49 166793	1.89028052
C21orf82 114036	1.890104644
FAM83H 286077	1.886659257
ZSWIM3 140831	1.884438466
ZNF583 147949	1.883249653
RAB6C 84084	1.881614787
ZNF3 7551	1.880918282
ZBTB34 403341	1.874917009
SYT15 83849	1.869268059
SLC3A2 6520	1.86864891
FBXL19 54620	1.867823382
KIF4B 285643	1.86727334
LMTK3 114783	1.867015804
AKT1S1 84335	1.863160087
CTDSP2 10106	1.861010066
JOSD2 126119	1.856359875
ZNF287 57336	1.854177649
ZNF417 147687	1.853174667
ADAT3 113179	1.851781544
ZNF625 90589	1.851629315
CAPN1 823	1.851567867
CCNI 10983	1.845867524
SPRED1 161742	1.844348157
TSP0 706	1.844334284
SAFB 6294	1.84189959
ZNF234 10780	1.841612423
NUDT16 131870	1.838551217
ADARB1 104	1.835131303
GJA5 2702	1.831568077

ZBTB1 22890	1. 827899728
LOC388955 388955	1. 826639691
KARS 3735	1. 826286938
ZNF225 7768	1. 826147238
FAM160A1 729830	1. 824835767
ZNF501 115560	1. 820375925
EXTL2 2135	1. 814530217
LOC100130331 100130331	1. 81246148
RGS7BP 401190	1. 808958895
GPR115 221393	1. 805111473
CEP68 23177	1. 80305719
SCAMP5 192683	1. 802100085
HDX 139324	1. 801001675
ARHGAP31 57514	1. 800063853
EPPK1 83481	1. 800032117
F12 2161	1. 799360532
ZNF132 7691	1. 798350588
? 653553	1. 797476297
ZNF627 199692	1. 796145631
SAMD9 54809	1. 791905967
FAT2 2196	1. 790731344
SH3GL1 6455	1. 789950929
NUDT22 84304	1. 788270129
SLC35A5 55032	1. 787227625
EIF5AL1 143244	1. 787005917
C6orf153 88745	1. 786942031
ZFP112 7771	1. 786844859
LOC344967 344967	1. 783969581
CLK2P 1197	1. 782095008
KIAA1549 57670	1. 781823729
MRPS10 55173	1. 781305683
TMEM136 219902	1. 781274238
FARSA 2193	1. 781233621
LYPLA2P1 653639	1. 780666384
ZNF468 90333	1. 77997949
LOC647288 647288	1. 777095643
OR1L8 138881	1. 775896297
TMEM102 284114	1. 773176534
ZNF630 57232	1. 773106328
RASL12 51285	1. 771639466
BAIAP2 10458	1. 766947665
NEK9 91754	1. 763790339
CLTB 1212	1. 759015142
AOC3 8639	1. 757889724

LHX5 64211	1. 754093801
FAM83B 222584	1. 753636981
SDC2 6383	1. 753308341
DAAM2 23500	1. 752178053
RPL13AP6 644511	1. 750895619
PRKD1 5587	1. 750837598
SMAD4 4089	1. 744820567
ZKSCAN4 387032	1. 744264967
SPTBN1 6711	1. 743521696
EIF3I 8668	1. 742412594
ARL4D 379	1. 742231468
CNOT8 9337	1. 741154443
MPDZ 8777	1. 738145189
PHRF1 57661	1. 738021629
CLIP1 6249	1. 737759546
ZNF785 146540	1. 736903362
GDPD1 284161	1. 736664254
C10orf140 387640	1. 735160691
PABPC3 5042	1. 735018152
YIPF2 78992	1. 734791876
EIF5A 1984	1. 733252105
RGAG4 340526	1. 733080688
CCDC124 115098	1. 732686109
RPSAP9 653162	1. 731715954
MYST4 23522	1. 731714586
SNURF 8926	1. 730092068
DOHH 83475	1. 729818129
DPYSL2 1808	1. 729781883
TAF1L 138474	1. 727927008
ZNF555 148254	1. 726990774
CLMN 79789	1. 72369895
MICALL1 85377	1. 723593249
PIK3R3 8503	1. 722976426
DISC1 27185	1. 722162614
EIF5B 9669	1. 721555597
IPPK 64768	1. 720912557
P704P 641455	1. 720561537
CDK5R1 8851	1. 718987198
RGS20 8601	1. 71802216
CC2D1A 54862	1. 71545273
PIGH 5283	1. 710363853
MAN2A2 4122	1. 706801938
EPC1 80314	1. 703911719
ZFPM1 161882	1. 703034345

HSD17B11 51170	1. 702248191
SCAF1 58506	1. 702098116
VEZF1 7716	1. 700394947
RAB2B 84932	1. 698529413
AHNAK2 113146	1. 698090859
ZNF154 7710	1. 698071482
S100A11 6282	1. 697647374
PLEKH02 80301	1. 697306238
RNF214 257160	1. 697261782
LOC729020 729020	1. 695998312
MORF4L1 10933	1. 6958068
DAPK3 1613	1. 695711059
BSN 8927	1. 695153882
NOP16 51491	1. 695073429
FEZ2 9637	1. 694124233
NCLN 56926	1. 693920943
CFL1 1072	1. 693482294
NFIB 4781	1. 693416184
SLC7A5 8140	1. 693082175
PPP1R13L 10848	1. 690688203
BAG3 9531	1. 69060956
HSPB1 3315	1. 689829893
CXADRP2 646243	1. 686503497
NR2C2 7182	1. 68590917
ZNF420 147923	1. 68578091
DDA1 79016	1. 6855767
SALL2 6297	1. 683553545
SLIT3 6586	1. 683550885
RHBDL2 54933	1. 682672919
MAGEH1 28986	1. 681298508
PRR5 55615	1. 680705128
MIB2 142678	1. 680568073
CTAGE1 64693	1. 680356105
ZNF45 7596	1. 677775073
TICAM1 148022	1. 677738462
FAM162B 221303	1. 677556542
RNASE1 6035	1. 677040026
SH2B3 10019	1. 676880226
ELAC1 55520	1. 676728426
TAS2R10 50839	1. 676660307
IRF6 3664	1. 676212583
KIAA1147 57189	1. 674543957
PEG3 5178	1. 673722962
CLEC2D 29121	1. 673164314

TMEM50B 757	1. 671360811
BICD2 23299	1. 668678172
SYT11 23208	1. 668472666
HIP1 3092	1. 668389721
RIN1 9610	1. 667536334
FGFBP1 9982	1. 667228342
SHE 126669	1. 666877364
USP51 158880	1. 665864321
ZNF704 619279	1. 665646991
CA5B 11238	1. 665363825
UBE3C 9690	1. 665049119
GALNT3 2591	1. 664014988
PTBP2 58155	1. 66285689
RECK 8434	1. 662545188
TNRC6C 57690	1. 661832803
ARSI 340075	1. 661162837
CXADRP3 440224	1. 66071827
GLCE 26035	1. 656793679
MRPL14 64928	1. 655823672
ZNF70 7621	1. 655208409
USP30 84749	1. 653181917
CDK14 5218	1. 652882161
AURKAIP1 54998	1. 652490093
SFN 2810	1. 651993102
ZNF24 7572	1. 651499515
UBE2M 9040	1. 651321927
DCLK2 166614	1. 65064144
ZNF566 84924	1. 650393457
ZNF252 286101	1. 650250171
ZNF738 148203	1. 649579286
FABP5 2171	1. 649506897
FAM161B 145483	1. 648285191
UBR7 55148	1. 648196247
CTBP2 1488	1. 647496442
MAP3K3 4215	1. 646830216
ZCCHC6 79670	1. 645851412
SERBP1 26135	1. 644376593
TOP1P1 7151	1. 642383102
TBC1D2B 23102	1. 642092081
RUNX1T1 862	1. 640344418
S100A16 140576	1. 640072021
LOC284441 284441	1. 63935719
ARNTL2 56938	1. 6391762
PURG 29942	1. 636910069

ZNF362 149076	1. 636814469
C19orf22 91300	1. 636376322
TPRXL 348825	1. 635720937
TRAK2 66008	1. 635212339
PYY2 23615	1. 634012225
NOC4L 79050	1. 633139521
RCBTB1 55213	1. 633108254
NAA15 80155	1. 63298905
ZNF443 10224	1. 632872265
ZNF230 7773	1. 632627432
GPR116 221395	1. 632382309
FAM178A 55719	1. 6312755
TRIM68 55128	1. 631182308
CMKLR1 1240	1. 630013843
UBE4A 9354	1. 629940828
OSTCL 202459	1. 629643931
RALBP1 10928	1. 62949566
FRY 10129	1. 629453877
ZNF187 7741	1. 629063113
GUK1 2987	1. 628396461
ZNF776 284309	1. 627631489
ZNF709 163051	1. 625806209
NBPF3 84224	1. 625323757
GRWD1 83743	1. 624173454
KPNA3 3839	1. 623743603
YWHAG 7532	1. 623097271
ATP5L2 267020	1. 622375899
KLHL15 80311	1. 621522466
ZC3H10 84872	1. 620627366
CDC14A 8556	1. 620359645
NSL1 25936	1. 620180801
ADRM1 11047	1. 620136123
ZNF821 55565	1. 619210466
HIST1H2BG 8339	1. 618836016
RPS6KA2 6196	1. 618710332
PLIN3 10226	1. 618023996
PROS1 5627	1. 617517616
CCDC86 79080	1. 617215663
ZIC5 85416	1. 617163178
LOC100271831 100271831	1. 615961451
BEST2 54831	1. 61541024
ARRB1 408	1. 615308499
PLSCR4 57088	1. 61440302
ENG 2022	1. 614168722

CTSL2 1515	1.612467128
HUS1B 135458	1.612401383
FAM71D 161142	1.612120983
FAT3 120114	1.611857158
HOMER1 9456	1.610928131
CSNK1A1 1452	1.610191518
ZNF689 115509	1.608474079
PSMB2 5690	1.608302563
ARHGEF35 445328	1.607277665
ARHGEF5 7984	1.606914042
JUP 3728	1.604825949
MVD 4597	1.604346755
DNAJB1 3337	1.60212967
TBC1D2 55357	1.601732715
PRKAG2 51422	1.601339841
TCEANC 170082	1.600553543
GNRHR 2798	1.600343595
TNFRSF19 55504	1.599415573
PLEKHN1 84069	1.598287639
ZNF32 7580	1.597027909
LOC643387 643387	1.596764311
TAF3 83860	1.595745969
AMMECR1 9949	1.595474681
TNKS1BP1 85456	1.595300063
SPATA7 55812	1.595122747
GMPPB 29925	1.594722946
DSP 1832	1.594475583

## I 与 II 间相关基因的假阳性检验结果

阈值	假发现率
0.1	0.321726
0.2	0.272677
0.3	0.22755
0.4	0.182128
0.5	0.131711
0.6	0.095917
0.7	0.06774
0.8	0.047991
0.9	0.032864
1	0.022049
1.1	0.014428
1.2	0.009329
1.3	0.00579

1.4	0.003636
1.5	0.002218
1.6	0.001171
1.7	0.000571
1.8	0.000259
1.9	0.000147
2	5.61E-07

## II 与III间相关基因

###probe	差异值
RPS6KA2 6196	2.167481027
C12orf39 80763	2.136203726
ELAVL3 1995	2.070371324
ADARB1 104	2.044578356
CYYR1 116159	2.04144345
FZD5 7855	2.040092562
SUSD2 56241	2.029957297
PLXND1 23129	2.0240167
PEAR1 375033	2.022406725
SLC9A3R2 9351	1.998222032
SDC2 6383	1.997235719
FAM155A 728215	1.991742573
EXTL2 2135	1.986774048
ADCY4 196883	1.976082783
SPRED2 200734	1.974724151
PITPNC1 26207	1.972207756
C20orf112 140688	1.930268283
LOC283267 283267	1.918775704
ITGA9 3680	1.916781901
JAM3 83700	1.916050314
C16orf52 730094	1.912918897
ZIK1 284307	1.912078572
DGKD 8527	1.911531619
CNRIP1 25927	1.909723054
FAM162B 221303	1.903730799
TAL1 6886	1.902887159
PURG 29942	1.900035994
OR51E2 81285	1.894643633
RASL12 51285	1.892438792
ZNF324B 388569	1.892431425
ECSCR 641700	1.887855415
NRIP2 83714	1.886486756



ITGB1 3688	1.883156521
FOXS1 2307	1.881612332
MMRN2 79812	1.876023328
PCDH12 51294	1.873618189
OR51E1 143503	1.871040013
CLMN 79789	1.871017423
CD34 947	1.864915782
KIAA1279 26128	1.859212892
HSPA12B 116835	1.856386072
ROBO4 54538	1.855774615
CTDSP1 58190	1.851721289
HYAL2 8692	1.850647272
GJA4 2701	1.840002556
RHOJ 57381	1.839650979
UBE2MP1 606551	1.834158521
RBMS2 5939	1.833409474
PROS1 5627	1.83314022
TRAK2 66008	1.822470797
EBF1 1879	1.821795788
S1PR1 1901	1.821679834
ACVRL1 94	1.818173349
SEPT4 5414	1.812167816
HECW2 57520	1.811907637
ARHGEF15 22899	1.811538174
NRGN 4900	1.810731994
RGS7BP 401190	1.809268469
KANK2 25959	1.805131254
TENC1 23371	1.804360557
DENND3 22898	1.801896772
TRPC4 7223	1.800864539
NUDT16 131870	1.797174296
ZNF30 90075	1.792898269
HEYL 26508	1.791475561
VANGL2 57216	1.79135521
CYGB 114757	1.785614978
DLC1 10395	1.782924053
CNTN4 152330	1.781609136
C10orf140 387640	1.779886168
APLNR 187	1.77775068
NFYB 4801	1.777204193
PARP1 142	1.774493891
CD93 22918	1.773302414
GFAP 2670	1.769463435
C3orf48 151649	1.766365709

TBXA2R 6915	1.765996576
SULT1C4 27233	1.765864153
ERG 2078	1.76484029
RGS5 8490	1.7612907
CLEC14A 161198	1.761006497
ZNF549 256051	1.758908679
FZD4 8322	1.758450139
ZNF98 148198	1.754660569
CXorf36 79742	1.754159119
GIMAP8 155038	1.753752069
TIE1 7075	1.751036052
CACNA1C 775	1.750868798
EPHA3 2042	1.74898264
C20orf160 140706	1.746328207
LHFP 10186	1.740323369
PLN 5350	1.738677168
PABPC5 140886	1.738210774
ESAM 90952	1.736688559
ARHGEF7 8874	1.735801263
CDH5 1003	1.732371418
FCN3 8547	1.732231755
GPR21 2844	1.731188237
DPYSL2 1808	1.728019017
PIPSL 266971	1.72393383
FBXL7 23194	1.723418013
SPTBN1 6711	1.721784803
ZNF449 203523	1.720970428
WFDC1 58189	1.719876336
PODXL 5420	1.719425516
AKAP2 11217	1.714204764
MYOCD 93649	1.712706425
PALM2-AKAP2 445815	1.712204362
SHE 126669	1.712042803
RAMP2 10266	1.711366701
TBX2 6909	1.710678835
GIPC3 126326	1.708456026
GRAP 10750	1.704887545
ZNF582 147948	1.702158126
KIAA1462 57608	1.700282344
LOC100132707 100132707	1.700098131
MYCT1 80177	1.698439934
PCDHA4 56144	1.696361862
ATP6V1G2 534	1.696051482
SPRY1 10252	1.695873334

PDE3A 5139	1.695106995
FAT2 2196	1.694909101
NOSTRIN 115677	1.692793436
NRXN2 9379	1.689680796
MGC16142 84849	1.689502591
PLSCR4 57088	1.688288092
SNAP25 6616	1.688244953
MRFAP1L1 114932	1.682003461
CUL1 8454	1.681888919
SLC22A10 387775	1.680435412
TGM2 7052	1.677495694
ARHGAP31 57514	1.677422199
ZNF418 147686	1.66814198
MSRB3 253827	1.667551807
TNS1 7145	1.667366532
FAM134A 79137	1.667094424
SLC44A5 204962	1.664599382
FGD5 152273	1.662670074
GRRP1 79927	1.661737301
PDE1C 5137	1.660388515
GPR116 221395	1.659710279
UBAP2L 9898	1.656172562
IFITM2 10581	1.654779916
RBM4B 83759	1.654329315
ENG 2022	1.654158814
FAM167B 84734	1.653820371
LDB2 9079	1.651351948
PCDH17 27253	1.650568164
HSPA4L 22824	1.64706646
SGCE 8910	1.645691993
SOX17 64321	1.645322137
SH2B3 10019	1.64490021
MGP 4256	1.642294547
CC2D1A 54862	1.642028229
SDC1 6382	1.642007565
GRK6 2870	1.641133007
SLC5A4 6527	1.639968686
LOC100134713 100134713	1.639493599
LYVE1 10894	1.635079187
ATP1B2 482	1.634525445
EXOC3L 283849	1.633445588
TEK 7010	1.631795399
TMEM88 92162	1.631238292
C18orf1 753	1.631010549

FAM177B 400823	1.629426515
PRKCD 5580	1.627676813
PLAC9 219348	1.625976626
RBPMS2 348093	1.624579154
CGRRF1 10668	1.624439397
MPRIP 23164	1.623962411
SH2D3C 10044	1.623090149
GIMAP6 474344	1.622969371
ARMCX1 51309	1.6225181
SHOX2 6474	1.622342317
ILK 3611	1.621324983
NOTCH4 4855	1.619995802
MCAM 4162	1.619589135
COPG2 26958	1.619417213
COL4A1 1282	1.619210547
CNP 1267	1.617327249
PIP5K1P1 206426	1.617270071
NRP1 8829	1.616839886
C22orf28 51493	1.616541912
IL18R1 8809	1.616361983
ZNF613 79898	1.615546485
DIXDC1 85458	1.611804588
GPR20 2843	1.611684695
ABI3BP 25890	1.611579512
PRCC 5546	1.609479992
CALCRL 10203	1.609259804
ZNF727 442319	1.608842169
GRB10 2887	1.60863907
LOC257358 257358	1.60862621
RNF125 54941	1.608499807
LMO2 4005	1.607812894

## II 与III间相关基因的假阳性检验结果

阈值	假发现率
0.1	0.332172
0.2	0.283415
0.3	0.237185
0.4	0.190528
0.5	0.141739
0.6	0.104761
0.7	0.074705
0.8	0.051685
0.9	0.034445

1	0.022169
1.1	0.013526
1.2	0.007681
1.3	0.004343
1.4	0.002207
1.5	0.001032
1.6	0.000459
1.7	0.000174
1.8	6.52E-05
1.9	2.99E-07
2	1.8E-09

### III与IV间相关基因

###probe	差异值
SLC9A3R2 9351	2.074022446
SELP 6403	1.846432012
EPAS1 2034	1.844245416
CDC42EP5 148170	1.825756648
RPS6KA2 6196	1.804247984
FCN3 8547	1.797922877
ITGA7 3679	1.792625419
EGFL7 51162	1.789000756
C13orf33 84935	1.786968126
TRIB2 28951	1.770836385
NPDC1 56654	1.736571515
JAM2 58494	1.730470342
MYOCD 93649	1.725972948
IL33 90865	1.712962515
C11orf70 85016	1.699292716
GATA2 2624	1.695242248
ZNF454 285676	1.694228737
CYYR1 116159	1.688493888
NRGN 4900	1.685258743
HYAL2 8692	1.682327129
NOS3 4846	1.680544776
FAM155A 728215	1.679705221
MAPK3 5595	1.679617927
EXOC3L2 90332	1.679219711
ZNF613 79898	1.674823236
IFITM2 10581	1.674499708
VWA1 64856	1.673602426
PI4KB 5298	1.671076093

GNA14 9630	1.669549458
ADCY4 196883	1.666984821
SHE 126669	1.663614973
RHD 6007	1.662539361
SLC02A1 6578	1.657584886
PITPNC1 26207	1.653031347
BMX 660	1.640305753
DGKD 8527	1.635385167
GRRP1 79927	1.632129516
SYT15 83849	1.629088878
HSPA12B 116835	1.628409931
PRTN3 5657	1.613610849
TCL6 27004	1.610245375
C3orf36 80111	1.605199593
KIAA0408 9729	1.602915655
TAL1 6886	1.599465913
APLNR 187	1.593967521
EFEMP1 2202	1.593310428
GJA5 2702	1.589494937
MMRN2 79812	1.58363549
ACVRL1 94	1.582707355
RAMP2 10266	1.575467762
ZIK1 284307	1.575148476
SPRY1 10252	1.574521101
SUGT1P1 441394	1.574137858
ZNF30 90075	1.570258523
LOC100134713 100134713	1.569760511
PEAR1 375033	1.567304826
LRRC32 2615	1.565912315
DARC 2532	1.565664748
GPR20 2843	1.559451995
CLEC14A 161198	1.557452531
PTGIS 5740	1.554928126
ENG 2022	1.554326183
PCDHB4 56131	1.548031228
DOCK9 23348	1.543107998
DHH 50846	1.538690832
CRIP2 1397	1.535763701
SLC22A4 6583	1.534617299
MS4A2 2206	1.534260199
C20orf160 140706	1.528470319
ARAP3 64411	1.527710546

### III与IV间相关基因的假阳性检验结果

阈值	假发现率
0.1	0.420267
0.2	0.361563
0.3	0.304263
0.4	0.2448
0.5	0.182929
0.6	0.131924
0.7	0.092196
0.8	0.061366
0.9	0.038773
1	0.023953
1.1	0.01421
1.2	0.008169
1.3	0.00442
1.4	0.002415
1.5	0.001259
1.6	0.000702
1.7	0.00029
1.8	0.000402
1.9	4.02E-06
2	8.03E-09